-21

Sequence

E710.00 101 A

PAT. 08/
3775888P

MW
9-27-94

PROKARYOTIC REVERSE TRANSCRIPTASE

RELATED CASES

Ins D1

~~This~~ which D1 is a continuation-in-part of prior copending U.S. patent application Serial No. 07/315,427, filed February 24, 1989 and since issued as U.S. Patent No. 5,079,151 on January 7, 1992, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/315,316, filed February 24, 1989 and since issued as U.S. Patent No. 5,320,958 on June 14, 1994, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/315,432, filed on February 24, 1989 and since abandoned, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/517,946, filed on May 2, 1990, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/518,749, filed on March 2, 1990, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/753,110, filed on August 30, 1991, which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/817,430, filed January 6, 1992, B1 which is a continuation-in-part of prior copending U.S. patent application Serial No. 07/979,447, filed November 20, 1992, respectively which are incorporated herein by reference.

Ins. B1

FIELD OF THE INVENTION

The invention relates to bacterial RT enzymes which are capable of synthesizing a hybrid RNA-DNA molecule, called msDNA together with the genes which synthesize the DNA and RNA portion of the molecule.

Another aspect of the invention relates to the isolation and purification of RTs from bacterium which is capable of synthesizing msDNA. The invention deals with groups of prokaryotes e.g., bacteria which are capable of synthesizing msDNAs by means of a reverse transcriptase. The bacterium capable of synthesizing msDNAs is identified by testing positive by an appropriate screening test.

This is the first time that, as taught in the subject parent patent applications, reverse transcriptase has been found and isolated from a prokaryote.

BACKGROUND OF THE INVENTION

Previously, there was described a chromosomal region of the bacterium Myxococcus xanthus which coded for the RNA and DNA portions of an msDNA. Dhundale et al. (Dhundale '87) "Structure of msDNA from Myxococcus xanthus: Evidence for a Long, Self-Annealing RNA precursor for the Covalently Linked, Branched RNA", Cell, Vol. 51, pages 1105-1112 (December 24, 1987). Dhundale et al. speculated that an Alu I nucleotide fragment contained all the essential coding regions to produce an msDNA. This speculation turned out to be in error.

The Alu I fragment of Dhundale et al., in fact, and inherently did not contain the gene sequence coding for an RT. The Alu I fragment was too short to code for the gene sequence coding for an RT. This was proven by way of sequence analysis by a computer program which searches for open reading frames that can potentially code for a protein. The print-out of the sequence analysis clearly shows that there is no translational reading frame in the Dhundale et al. fragment open across a stretch of DNA sufficiently long enough to encode any reverse transcriptase.

What is reported in Dhundale et al. in 1987 with respect to a bacterial reverse transcriptase was totally contrary to accepted dogma at that time about the distribution of these enzymes, i.e., that they were present only in viruses which infect eukaryotic organisms.

For the 20 years since the discovery of reverse transcriptase, it was believed that these enzymes were restricted to viruses which infect eukaryotic cells. Now, in accordance with the invention, reverse transcriptases have been identified in bacteria.

SUMMARY OF THE INVENTION

In accordance with the invention, it is shown that various bacteria have nucleotide sequences named "retrons" which encode reverse transcriptases (RTs) which are capable of synthesizing msDNAs. The invention also relates to the isolated and purified bacterial RTs. It has also been determined that the RTs of the bacteria which synthesize msDNAs possess common conserved nucleotide sequences and amino acid residues.

Representative members of the Enterobacteriaceae, Rhizobiaceae and Mycobacteriaceae families are demonstrated to be capable of synthesizing msDNA. These bacteria can be screened for the capability of synthesizing msDNA by an RT labeling or extension in vitro test.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows the restriction map of the 3.4 kb fragment around msd and downstream of msr.

Figure 2 shows the nucleotide sequence of the chromosomal region encompassing the msDNA and msd RNA coding regions and an ORF region downstream of msr and the amino acid sequence of Mx162-RT.

Figure 3 shows the amino acid sequence alignment of the msDNA-Mx162 ORF with a portion of the retroviral Pol sequences from HIV and HTLV1 and the ORF of msDNA-Ec67.

Figure 4 shows the sequence similarity of the msDNA-Mx162 reverse transcriptase with other retroelements.

Figure 5 shows the sequence comparison of the regions around the YXDD box of various reverse transcriptases.

Figure 6 shows the detection of msDNA in a clinical isolate of E. coli.

Figure 7 shows the complete primary and proposed secondary structure of msDNA-Ec67.

Figure 8 shows the determination of the RNA nucleotide sequence for the branched RNA linked to msDNA.

Figure 9 shows the southern blot analysis of E. coli Cl-1 Chromosomal DNA(A) and analysis of msDNA synthesis by pCl-1E and pCl-1P(B).

Figure 10 shows the restriction map of the 11.6 kb Eco RI fragment.

Figure 11 shows the nucleotide sequence of the region from the E. coli Cl-1 chromosome encompassing the msDNA, msd RNA and ORF coding regions and the amino acid sequence of Ec67-RT.

Figure 12 shows the amino acid sequence alignment of the E. coli msDNA ORF with a portion of the retroviral Pol sequence from HIV and HTLV1.

Figure 13 shows the detection of RT activity from various cell extracts.

Figure 14 shows the amino acid sequence alignment of bacterial RTs.

Figure 15 shows the nucleotide and amino acid sequence of Mx65-RT.

Figure 16 shows the nucleotide and amino acid sequence of Sa163-RT.

Figure 17 shows the nucleotide and amino acid sequence of Ec73-RT.

Figure 18 shows the nucleotide and amino acid sequence of Ec86-RT.

Figure 19 shows the nucleotide and amino acid sequence of Ec107-RT.

Figure 20 shows the msDNAs from total RNA prepared from each bacterial strain were specifically labeled with ^{32}P by the RT extension method (12, 14).

Figure 21 shows a collection of 63 rhizobial isolates screened for the presence of msDNA by the RT extension method.

DETAILED DESCRIPTION OF THE DRAWINGS

Figure 1. Restriction Map of the 3.4-kb fragment Around msd and Downstream of msr.

The locations and the orientation of msDNA and msdRNA are indicated by a small arrow and an open arrow, respectively. A large solid arrow represents an ORF and its orientation. The only two AluI sites (A and B) are shown and the DNA sequence between AluI (A) and AluI (B) was determined previously by Yee et al. (1984).

Figure 2. Nucleotide Sequence of the Chromosomal Region Encompassing the msDNA and msdRNA Coding Regions and an ORF Region Downstream of msr Seq. ID NO. 1 and SEQ ID NO:2.

The upper strand beginning at the Alu I (A) site (see Figure 1) and ending just beyond the ORF is shown. Only a part of the complementary lower strand is shown from base-301 to -600. The boxed region of the upper strand (332-408) and the boxed region of the lower strand (401-562) correspond to the sequences of msdRNA and msDNA respectively (Dhundale et al., 1987). The starting sites for DNA and RNA and the 5' to 3' orientations are indicated by open arrows. The msdRNA and msDNA regions overlap at their 3' ends by 8 bases. The circled G residue at position 351 represents the branched rG of RNA linked to the 5' end of the DNA strand in msDNA. Long solid arrows labeled a1 and a2 represent inverted repeat sequences proposed to be important in the secondary structure of the primary RNA transcript involved in the synthesis of msDNA (Dhundale et al., 1987). The ORF begins with the initiation codon at base 640. Single letter designations are given for amino acids. The YXDD amino acid sequence highly conserved among known RT proteins is boxed. Numbers on the right hand column enumerate the nucleotide bases and numbers with a* enumerate amino acids. Small vertical arrows labeled Alu I and SmaI locate the Alu I and SmaI restriction cleavage sites, respectively. The DNA sequence was determined by the chain termination method (Sanger et al., 1977) using synthetic oligonucleotides as primer.

Figure 3. Amino acid Sequence Alignment of the msDNA-Mx162 ORF Seq. ID NO. ~~4~~ 5 with a Portion of the Retroviral Pol Sequences from HIV Seq. ID. NO: ~~2~~ 3 and HTLV1 Seq. ID. NO ~~3~~ 4 and the ORF of msDNA-Ec67 Seq. ID NO. ~~5~~ 6.

Amino acid sequences are compared with matching residues assigned as follows: (o) amino acid residues shared by all four proteins; (o) amino acid residues shared by msDNA-Mx162 and msDNA-Ec67 RTs; (x) amino acid residues shared by msDNA-Mx162 RT with HIV or HTLV1 RTs. Amino acid sequences showed are from residue-177 to -439 for HIV RT (Ratner et al., 1985); residue-15 to -277 for HTLV1 RT (Seiki et al., 1983); residue-32 to -291 for Ec-67 RT (Lampson et al., 1989); and residue-170 to -435 for Mx-162 RT (this work). The YXDD consensus sequence is outlined with a box.

Figure 4. Sequence Similarity of the msDNA-Mx162 Reverse Transcriptase with Other Retroelements. A. Sequence similarity of the region from residue-18 to -128 of the msDNA-Mx162 RT (see Figure 2) Seq. ID NO. ~~6~~ 7 with a carboxyl terminal region of integrase of Moloney murine leukemia virus (Mo-MLV) (residue-1070 to -1179; Shinnick et al., 1981) Seq. ID NO. ~~7~~ 8. B. Comparison of the sequence from residue-411 to -485 of the msDNA-Mx162 RT (see Figure 2) Seq. ID NO. ~~8~~ 9 with the sequence from residue-396 to -461 of the gap protein of human immunodeficiency virus (HIV; Ratner et al., 1985) Seq. ID NO. ~~9~~ 10.

Figure 5. Sequence Comparison of the Regions Around the YXDD Box of Various Reverse Transcriptases.

The region from residue-304 to residue-371 of the msDNA-Mx162 RT (see Figure 2) is aligned with various RTs from different sources. The identical amino acid residues with the msDNA-Mx162 RT are indicated by open circles. The YXDD sequences are boxed. The residue numbers for the amino terminal residues and for the carboxyl terminal residues are indicated by the left and the right hand sides of the sequences, respectively. Mx-162 RT from this work (Figure 2) Seq. ID No. ~~10~~ 11; Ec-67 RT from Lampson et al. (1989) Seq. ID No. ~~11~~ 12; Ec-86 RT from Lim and Maas (1989) Seq. ID. No. ~~12~~ 13; HIV RT from Ratner et al. (1985) Seq. ID No. ~~13~~ 14; HTLV1 RT from Seiki et al. (1983) Seq. ID No. ~~14~~ 15; Mo-MLV RT from Shinnick et al. (1981) Seq. ID NO. ~~15~~ 16; RSV (Rous sarcoma virus) RT from Dickson et al. (1982) Seq. ID NO. ~~16~~ 17; BLV (bovine leukemia virus) RT from Rice

GB et al. (1985), Seq. ID NO. ~~17~~ 18; Mt. plasmid (Neurospora mitochondrial plasmid) RT from Nargang et al. (1984), Seq. ID. No. ~~18~~ 19; 17.6 GB Drosophila retrotransposon from Saigo et al. (1984), Seq. ID NO. ~~19~~ 20; gypsy Drosophila retrotransposon from Yuki et GB al. (1986), Seq. ID. NO. ~~20~~ 21; Tal-3 plant (Arabidopsis thaliana) retrotransposon from Voytas and Ausubel (1988), Seq. ID NO. ~~22~~ 23; and GB Ty912 yeast retrotransposon from Clare and Farabaugh (1985), Seq. ID NO. ~~23~~ 24. Small arrows in Copia, Seq. ID NO. ~~21~~ 22, Tal-3 and Ty912 indicate positions of insertions of extra sequences of 18, 18 and 13 residues, respectively. B, Phylogenetic relationships among various RTs listed in A. The branching positions are arbitrarily illustrated.

Figure 6. Detection of msDNA in a clinical isolate of E. coli. Total RNA, prepared (Maniatis et al., 1982) from a 5-ml culture, was added to 50 μl of a reaction mixture containing: 50 mM Tris-HCl (pH8.3); 6 mM $MgCl_2$; 40 mM KCl; 5 mM DTT; 1 μM dATP, dTTP, and dGTP; 0.04 μM dCTP; 0.2 μM [α-^{32}P]dCTP; and 10 units of AMV-RT (Boehringer Mannheim). The reaction mixture was incubated at 37°C for 30 min. followed by extraction with 50 μl phenol-chloroform (1:1) and ethanol precipitation. The samples were electrophoresed on a 4% acrylamide - 8 M urea gel. Lanes: (S) molecular weight markers; MspI digest of pBR322 end-labeled with [α-^{32}P]dCTP and the Klenow fragment of DNA polymerase I, (1) E. coli K-12 strain C600, (2) the same as in lane 1 except the sample was treated with RNase A (5 μg, 10 min at 37 °C) just prior to electrophoresis, (3) clinical isolate Cl-1, (4) clinical isolate Cl-1 treated with RNase A. The clinical isolate was identified as Escherichia coli (The clinical E. coli strains were urinary tract isolates kindly provided by Dr. Melvin Weinstein from the microbiology laboratory, R.W. Johnson Hospital, New Brunswick, NJ. The clinical strain Cl-1 was identified using the API-20E identification system (API laboratory products) and gave a typical E. coli profile number of 5044552).

Figure 7. The complete primary and proposed secondary structure of msDNA-GB Ec67. The DNA sequence, Seq. ID NO. ~~25~~ 26, was determined by the Maxam and Gilbert method (Maxam et al., 1980) using 3'-end labeled msDNA. The RNA sequence, Seq. ID NO 24, 25 (msdRNA; boxed region) was determined using base-specific RNases as previously described (Dhundale et al., 1987). The 2',5' Branched linkage

between the 15th rG residue and the 5' end of the DNA strand was determined using the debranching enzyme from HeLa cells as described previously (Dhundale et al., 1987; Furuichi et al., 1987; Ruskin et al., 1985; Arenas et al., 1987; the debranching enzyme was a gift from Jerard Hurwitz). The branched rG at position 15 is circled, and both RNA and DNA are numbered from their 5' ends.

Figure 8. Determination of the RNA nucleotide sequence for the branched RNA linked to msDNA. Total RNA was prepared from the clinical strain Cl-1 and fractionated on a 5% acrylamide gel. msDNA containing full length RNA was eluted from the gel. This fraction was then labeled at the 5' end of the RNA (piece 8, SEQ ID NO:25) with [γ-^{32}P]ATP and T4 polynucleotide kinase. The 5' end labeled RNA linked to msDNA was again purified on an 18% acrylamide - 8M urea sequencing gel. The labeled RNA was then sequenced using limited digestion with base-specific RNases as described previously (Dhundale et al., 1987). Lanes: OH$^-$, partial alkaline hydrolysis ladder; (0.5 M sodium bicarbonate/carbonate pH9.2); -E, no enzyme treatment of the labeled RNA linked to msDNA; T1, RNase T1 (1U/reaction, 55°, 15 min.); U2, RNase U2 (1U and 0.5U/reaction, 55°, 15 min.); PhyM, RNase PhyM (1U/reaction, 55°, 15 min); Bc, RNase B. cerus (2U/reaction, 55°, 15 min.); CL3, RNase CL3 (2U/reaction, 37°, 15 min.). The large gap in the sequence gel is due to msDNA linked at the rG residue at position 15 by a 2',5' phosphodiester linkage (Furuichi et al., 1987). The RNA sequence at the 3'-end region from the branched rG residue (the upper part of the gel) was determined from 6% gel (data not shown).

Figure 9. Southern blot analysis of E. coli Cl-1 chromosomal DNA(A) and analysis of msDNA synthesis by pLl-1E and pCl-1P(B). A: The chromosomal DNA was digested with EcoRI (lane 1), HindIII (lane 2), BamHI (lane 3), PstI (lane 4), and BglII (lane 5). For each lane, 3 μg of the DNA digest was applied to a 0.7% agarose gel. After electrophoresis the gel was blotted to a nitrocellulose filter, and hybridization analysis was carried out according to Southern (Southern, 1975) using msDNA labeled by AMV-RT with [α-^{32}P]dCTP as a probe. Numbers at the left represent the molecular weights in kb. B: Total DNA prepared from each strain was treated with RNase A,

separated on a 5% acrylamide gel and stained with ethidium bromide. Lane S, pBR322 digested with MspI used for molecular size markers; lane 1, DNA prepared from the host strain CL-83(recA⁻); lane 2, CL-83 (recA⁻) transformed with plasmid pCl-1E (11.6 kb EcoRI fragment; see Figure 5); lane 3, with plasmid pCl-1P (2.8-kb PstI(a)-PstI(b) fragment; see Figure 5). An arrow indicates the position of msDNA.

Figure 10. Restriction map of the 11.6-kb EcoRI fragment. In the Cl-1E map, the left-hand half (EcoRI to HindIII) was not mapped. In the Cl-1EP5 map, the locations and the orientations of msDNA and msdRNA are indicated by a small arrow and an open arrow, respectively. A large solid arrow represents an ORF and its orientation.

Figure 11. Nucleotide sequence of the region from the E. coli Cl-1 chromosome encompassing the msDNA and msdRNA coding regions and an ORF downstream of the msdRNA region Seq. ID No: ~~36~~27. The entire upper strand beginning at the BalI site (see Figure 5) and ending just beyond the ORF is shown. Only a part of the complementary lower strand is shown from base 241 to 420. The long boxed region of the upper strand (249-306) corresponds to the sequence of the branched RNA (msdRNA; see Figure 7) portion of the msDNA molecule. The boxed region of the lower strand corresponds to the sequence of the DNA portion of msDNA (see Figure 7). The starting site for DNA and RNA and the 5' to 3' orientations are indicated by large open arrows. The msdRNA and msDNA regions overlap at their 3' ends by 7 bases. The circled G residue at position 263 represents the branched rG of RNA linked to the 5' end of the DNA strand in msDNA. Long solid arrows labeled a1 and a2 represent inverted repeat sequences proposed to be important in the secondary structure of the primary RNA transcript involved in the synthesis of msDNA (Dhundale et al., 1987). Note that the nucleotide at position 257 (U on the RNA transcript) and the nucleotide at position 373 (G on the RNA transcript) form a U-G pair in the stem between sequence a1 and a2. The proposed promoter elements (-10 and -35 regions) for the primary RNA transcript are also boxed. The ORF begins with the initiation codon at base 418. Single letter designations are given for amino acids. The YXDD

amino acid sequence conserved among known RT proteins is boxed. Numbers on the right hand column enumerate the nucleotide bases and numbers with a* enumerate amino acids. Small vertical arrows labeled H and P locate the HindIII and PstI restriction cleavage sites, respectively. The DNA sequence was determined by the chain termination method (Sanger et al., 1977) using synthetic oligonucleotides as primers.

G B

G B

Figure 12. Amino acid sequence alignment of the E. coli msDNA Seq ID No. ~~29~~ 31 ORF with a portion of the retroviral Pol sequence from HIV, Seq. ID No. ~~27~~ 29, and HTLV1, Seq. ID No. ~~28~~ 30~~.~~ Amino acid sequences are compared with matching residues assigned as follows: (+) amino acid common to msDNA and HIV RTs; (o) amino acid shared by msDNA and HTLV1 RTs; and (o) amino acid shared by all three proteins. Arrows divide the protein sequences into three functional domains (Toh et al., 1983; Geng et al., 1985; Varmus, 1985, Tanese et al., 1988): An amino terminal RT domain, a carboxy terminal RNase H region, and a central "tether" region. The specific amino acid residues for the RT, tether, and RNase H domains, for each protein are: HIV, 177-439, 440-600, 601-722 respectively; HTLV1, 15-277, 278-462, 463-592 respectively; msDNA ORF, 32-290, 291-465, 466-586 respectively. The YXDD polymerase consensus sequence is outlined with a box.

Figure 13. Detection of RT activity from various cell extracts. Crude cell extracts were prepared from E. coli strain C2110 (polA$^-$) (Tanese et al., 1985; Tanese et al., 1986. E. coli strain C2110 (polA1$^-$) was a gift from M. Roth and S. Goff) containing plasmid pCl-1EP5 encoding the msDNA-ORF (see Figure 10) as well as the vector plasmid (pUC9; Yanisch-Perron et al., 1985) alone. Extracts were also prepared from the E. coli strain PRTS7-1 (polA+) containing the cloned M-MuLV RT gene (Varmus et al., 1985; Tanese et al., 1977; Tanese et al., 1985; Tanese et al., 1986. Crude extracts were prepared essentially as described (Roth et al., 1985; Hizi et al., 1988). Crude extract equivalent to 15 μg total protein was added to a 50 μl reaction cocktail (50 mM tris-HCl pH7.8, 10 mM DTT, 60 mM NaCl, 0.05% NP-40, 10 mM $MgCl_2$, 0.5 μg poly(rC)-oligo(dG), and 0.1 μM [α-^{32}P]dGTP and incubated at 37°C for one hour. Five μl of the reaction mixture was then spotted onto

DEAE paper (DE81; Whatman Inc.). The paper was washed to remove unincorporated label (Tanese et al., 1985; Tanese et al., 1986) and then exposed to an X-ray film. In row (A) all reactions contain added template primer (poly rC-dG). Row (B) contains control reactions in which no template-primer is added. Columns contain the designated cell extracts: M-MuLV, cloned Moloney Murine Leukemia Virus RT gene; pGB2 (Churchward et al., 1984), vector plasmid in strain C2110; pCl-1EP5, recombinant plasmid with the cloned msDNA gene. The large amount of background activity observed with the M-MuLV control extract is due to the activity of DNA Polymerase I since this extract is obtained from a PolA$^+$ strain (HB101).

Figure 14 shows the amino acid sequence alignment of bacterial RT carried out according to Xiong and Eickbush (1990). Amino acids highly conserved in eukaryotic RTs are shown at the top of the sequences. These amino acids include largely unvaried residues or chemically similar residues. (h) Hydrophobic residue; (p) small polar residues; (c) charged residue. Amino acids conserved in all seven bacterial RTs (identical residues plus functional conserved residues indicated by h for hydrophobic residues or p for polar residues) are marked by solid dots at the bottom of the sequences. The consensus sequence shown at the bottom of the sequences is determined when five out of seven sequences contain an identical or a chemically similar residue (h, hydrophobic residue; p, charged and polar residue). The subdomains 1 to 7 are according to Xiong and Eickbush (1990), which are boxed and indicated by numbers. The highly conserved YXDD sequences are also boxed. Numbers on the right indicate the amino acid positions from the amino terminus for each RT. Sources for the sequences are Sal63 (Hsu et al. 1992b) Seq. ID No. ~~30~~ 32, Mx162 (Inouye et al. 1989) Seq. ID. No. ~~31~~ 33, Mx65 (Inouye et al. 1990) Seq. ID No. ~~32~~ 34, Ec67 (Lampson et al. 1989b) Seq. ID No ~~33~~ 35, Ec86 (Lim and Maas 1989) Seq. ID No 36, Ec73 (Sun et al. 1991) Seq. ID No 37 ~~35~~, and Ec107 (Herzer et al. 1992) Seq. ID. No. ~~36~~ 38.

Figure 15 shows nucleotide sequence of the chromosomal region encompassing the Mx65-msDNA and msdRNA coding regions and an ORF region downstream of msr Seq. ID No. ~~37,~~ 40. The sequence covers from the Alu I(A) site to 78 bp downstream of the ORF. The complementary strand is only shown from bases 121-300. The boxed region of the upper strand (positions 143-191) and the boxed region of the lower strand (positions 186-250) correspond to the sequences of msdRNA and msDNA, respectively. The starting sites for DNA and RNA and the 5' to 3' orientation are indicated by open arrows. The msdRNA and msDNA regions overlap at their 3' ends by 6 bases. The circled G residue at position 206 represents the branched guanosine of RNA linked to the 5' end of the DNA strand in msDNA. Long solid arrows labeled a1 and a2 represent inverted repeat sequences proposed to be important in the secondary structure of the primary RNA transcript involved in the synthesis of msDNA. The ORF begins with the initiation codon at base 279. The YXDD amino acid sequence highly conserved among known RT proteins is boxed. Numbers on the right-hand column enumerate the nucleotide bases, and numbers with asterisks enumerate amino acids (single-letter code). The DNA sequence was determined by the chain-termination method using synthetic oligonucleotides as primers.

Figure 16 shows nucleotide sequences of 3,060 bases encompassing msr, msd, and the RT gene of S. aurantiaca. Seq. ID No ~~37~~ ~~38~~ 42 The sequence from base 421 to base 720 which contains msr and msd is shown double stranded. The boxed regions of the upper strand (bases 440 to 540) and the lower strand (bases 508 to 670) correspond to the sequences of msdRNA and msDNA, respectively. The starting sites for msDNA and msdRNA are indicated by open arrows. The circled G at the position 458 is the branched rG of msdRNA linked to the 5' end of msDNA. Long solid arrows labeled with a1 and a2 represent inverted repeated sequences proposed to form the secondary structure in the primary RNA transcript which serves to prime msDNA synthesis. Amino acids are indicated by single letters. The YXDD sequence highly conserved among known RTs is boxed. X^e and B^f sites are indicated by arrows. Numbers on the right-hand side and numbers with asterisks represent numbers for bases and amino acids, respectively.

Figure 17 shows the sequences of msdRNA and msDNA SEQ ID NO:43, SEQ ID NO:44 and SEQ ID NO:45 ~~Seq. ID No. 39~~ which are boxed and their orientations are indicated by open arrows. The branched G residue at position 10425 is circled. The inverted repeat sequences require for the biosynthesis of msDNA - Ec73 are shown by arrows labeled a1 and a2. Amino acid residues of Ec73-RT are shown by a single-letter code put at the center of each codon. ~~Seq ID NO. 40,~~

Figure 18 shows the restriction map of the 3.5 kb insert of pDB808 and nucleotide sequence of chromosomal determinants of the msDNA-RNA compound of E. coli B. (A) Restriction map of the 3.5 kb insert of clone pDB808. The solid bar represents the region whose sequence is presented in (B). Transcription is from left to right. Restriction enzymes are: P, Pstl, H, Hpal; B, Bglll; X, Xhol. (B) Nucleotide sequences of the chromosomal determinants. Seq. ID No. ~~44~~ 46, and SEQ ID NO:47 Only the strand corresponding to the transcript is shown. Nucleotides are numbered starting from the first base observed in the msdRNA. The mdsRNA coding region is overlined, and the msDNA coding region is underlined. The msDNA sequence is complementary to the sequence shown in this figure. Inverted repeats are indicated by double-dashed lines. The G at position 14 is the branched guanylate of msdRNA in the msDNA-RNA compound. IR, 12 bp inverted repeat.

Figure 19 shows sequence of the retron and flanking regions of Ec107. Seq. ID NO. ~~42.~~ 48 and SEQ ID NO:49 The sequences corresponding to the K-12 genomic DNA are shown in lower case letters from bases 1-99 and 1400-1540. The msRNA and msDNA regions are boxed. Also indicated are the a1-a2 conserved inverted repeats (indicated by arrows) and the branched G, which is circled. The RT consists of 319 amino acids and contains the YXDD sequence (boxed) which is highly conserved among known RTs. The transcription start site occurs at base 170; a possible terminator is indicated by head-to-head arrows following the RT coding region. Primer extension was utilized in order to determine the transcription start site. These sequence data will appear in the EMBL/GenBank/DDJB Nucleotide Sequence Data Libraries under the accession number X62583.

DETAILED DESCRIPTION OF THE INVENTION

The description which follows describes msDNA and RT from Myxococcus xanthus. This is a typical bacterium which belongs to a genus of bacteria, whose representative members possess an RT capable of synthesizing msDNA.

The existence of a peculiar branched RNA-linked DNA molecule called msDNA (multicopy single-stranded) has been demonstrated in various myxobacteria, Gram-negative soil bacteria (Yee et al., 1984; Dhundale et al., 1985; Furuichi et al., 1987a,b; Dhundale et al., 1987; Dhundale et al., 1988b). msDNA (msDNA-Mx162) from Myxococcus xanthus consists of 162-base single stranded DNA, the 5' end of which is linked to the 2' position of the 20th rG residue of a 77-base RNA molecule (msdRNA) by a 2', 5'-phosphodiester linkage (Dhundale et al., 1987). It exists at a level of approximately 700 copies per genome. Stigmatella aurantiaca also possesses an msDNA (msDNA-Sal63) which is highly homologous to msDNA-Mx162 (Furuichi et al., 1987b). In addition to msDNA-Mx162, M. xanthus has another smaller species of msDNA (mrDNA or msDNA-Mx65), which has no primary sequence homology with msDNA-Mx162 or msDNA-Sal63 (Dhundale et al., 1988b). However, all msDNAs so far characterized share key structural features such as a branched rG residue, stem-and-loop structures in RNA and DNA molecules, and a DNA-RNA hybrid at the 3' ends of DNA and RNA molecules.

Previously it was predicted that reverse transcriptase is required for msDNA biosynthesis on the basis of the finding that msdRNA is derived from a much longer precursor, which can form a very stable stem-and-loop structure (Dhundale et al., 1987). This precursor molecule was proposed to serve as a primer for initiating msDNA synthesis as well as a template to form the branched RNA-linked msDNA. The latter reaction requires reverse transcriptase activity. In M. xanthus, the region coding for the RNA molecule (msr) is located on the chromosome in the opposite orientation to the msDNA coding region (msd) with the 3' ends overlapping by 6 bases for msDNA-Mx65 (Dhundale et al., 1988b) or by 8 bases for msDNA-Mx162 (Dhundale et al., 1987). In addition, as in all the msDNAs found in myxobacteria, there is an inverted repeat comprised of a 14-base sequence for msDNA-Mx65 (Dhundale et al., 1988b) or a 34-base sequence for msDNA-Mx162 (Dhundale et al., 1987) and a 33-base sequence for msDNA-Sal63 (Furuichi et al., 1987b) immediately upstream of the branched G residue and a sequence immediately upstream of the msDNA coding region. As a result of this inverted repeat, a longer primary transcript beginning upstream of the RNA coding region and extending through the msDNA coding region is considered to self-anneal and form a stable secondary structure. When three base mismatches were introduced into the secondary structure immediately upstream of the branched rG residue, msDNA synthesis was almost completely blocked. However, if three additional base substitutions were made on the other strand to resume the complementary base pairing, msDNA production was restored (Hsu et al., 1989). This result strongly supports the proposed model for msDNA synthesis.

It was also shown that a deletion mutation at the region 100 base pairs (bp) upstream of the DNA coding region (msd) and an insertion mutation at a site 500 bp upstream of msd caused a significant reduction in msDNA production (Dhundale et al., 1988a). This indicates that there is a cis- or trans-acting positive element required for msDNA synthesis in this region. In this report we determined the DNA sequence of this region and found an opening reading frame (ORF) coding for 485 amino acid residues beginning with an initiation codon, ATG, which is located 77 bp upstream of msd (or 231 bp downstream of msr). The very close proximity between msd and the ORF suggests that they may be transcribed as a single transcript. The amino acid sequence of the ORF shows similarity with retroviral reverse transcriptases. We discuss a possible origin of the reverse transcriptase gene as well as a possible relationship between the msDNA system and retroviruses. Recently, some strains of Escherichia coli were found to produce msDNA and the gene for reverse transcriptase which is essential for msDNA production, is linked to the msd region, (Lim and Maas, 1989; Lampson et al., 1989b). Comparison of the msDNA systems of M. xanthus and E. coli raises an intriguing question as to how the extensive diversity found in msDNA systems has emerged in bacteria and what possible functions msDNA may have.

In a preceding paper, it was demonstrated that msDNA is in fact synthesized by reverse transcriptase in a cell-free system in M. xanthus (Lampson et al., 1989a).

Reverse transcriptases are isolated, and if desired, purified, and biological characterization carried out, if desired, by known methods such as those described in Lampson, B.C., M. Viswanathan, M. Inouye and S. Inouye, "Reverse Transcriptase from *Escherichia coli* Exists as a Complex with msDNA and is Able to Synthesize Double-stranded DNA", *J. Biol. Chem.* 265: 8490-8496 (1990), which is incorporated by reference as if fully set forth herein.

RESULTS AND DISCUSSION

Identification of an ORF Associated with *msd*

On the basis of mutations closely associated with *msd* which significantly reduce msDNA production, it was assumed that in this region there is a cis- or trans-acting element which is essential for msDNA synthesis (Dhundale *et al*., 1988a). Figure 1 shows a restriction map around *msd*. The msDNA coding region is shown by a thin arrow from right to left (*msd*), and the msdRNA coding region by a thick open arrow (*msr*). In the previous work (Dhundale *et al*., 1988a), two mutations were constructed; one, a deletion mutation in which the sequence from *Alu I*(b) to SmaI was replaced by a gene for kanamycin resistance (see Figure 1), and the other an insertion mutation at the SmaI site by a gene for kanamycin resistance (see Figure 1).

In order to elucidate the properties of the element required for msDNA production, the DNA sequence of the region upstream of *msd* was determined as shown in Figure 2. A long open reading frame (ORF) beginning with an initiation codon was found 77 bases upstream of *msd*. The ORF is preceded by a ribosome binding sequence of AGG (residue 630 to 632) 7 bases upstream of the initiation codon. The ORF codes for a polypeptide of 485 amino acid residues. The *Alu I*(b) and SmaI sites (see Figure 1), where mutations inhibiting msDNA synthesis were created, are located at amino acid residue-12 and -142 of the ORF, respectively or at the nucleotide sequence from residue -672 to -675, and from residue-1061 to -1066, respectively (Figure 2). In Figure 2, *msd* or the DNA sequence corresponding to the msDNA sequence is indicated by the closed box on the lower strand and the orientation is from right to left. Similarly, the msdRNA sequence (*msr*) is also indicated by the closed box on the upper strand and the orientation is from left to right. The msd and msr regions overlap by 8 bases. An inverted repeat is also indicated by arrows with letters a1 and a2. This inverted repeat comprises a 34-base sequence immediately upstream of the branched G residue (residue 317 to 350; sequence a2 in Figure 2) and another 34-base sequence at the 3' end (residue 597 to 564; sequence a1). This inverted repeat is essential to form a stem structure which provides a stable secondary structure in a long primary transcript. This secondary structure is considered to serve as the primer as well as the template for msDNA synthesis (Dhundale et al., 1987; Hsu et al., 1989).

Sequence Similarity with Retroviral Reverse Transcriptases

When the amino acid sequence of the ORF was compared with known proteins, a striking similarity was found between the sequence from Leu-308 to Ser-351 and retroviral reverse transcriptases (RT). In particular, this region contains the YXDD sequence, the highly conserved sequence in all known RTs. This sequence (Tyr-344 to Asp-347) is boxed in Figure 2. In Figure 3, the ORF sequence of 266 amino acid residues from Ala-170 to Lys-435 is compared with RTs from HIV (human immunodeficiency virus; Ratner et al., 1986) and HTLV1 (human T-cell leukemia virus type 1; Seiki et al., 1983). As mentioned above, within the sequence of 44 amino residues from Leu-308 to Ser-351, there are 14 and 12 identical residues with HIV (32%) and HTLV1 (27%), respectively. The entire RT domains of HIV and HTLV can also be aligned with the ORF sequence from Ala-170 to Lys-435, with much less similarity as shown in Figure 3. However, the same region was found to be extremely well aligned with the RT which was recently found in a clinical strain of *Escherichia coli* (Lampson et al., 1989b). This *E. coli* RT consists of 586 amino acid residues, and its amino terminal domain (residue-32 to -291) and the carboxyl terminal domain (residue-466 and -586) have been demonstrated to have sequence similarity with retroviral RT and ribonuclease H. This RT gene from *E. coli* was shown to be required for the production of msDNA (msDNA-Ec67) and to have reverse transcriptase activity (Lampson et al., 1989b). Figure 3 shows that the sequence similarity between *E. coli* and *M. xanthus* RTs is distributed within almost the entire RT region; in particular in the region from Tyr-181 to Ser-212, 15 out of 32 residues are identical (47% similarity);

in the region from Gly-226 to Gly-265, 19 out of 40 residues (48% similarity); in the region from Leu-308 to Ser-351, 26 out of 44 residues (59% similarity); and in the region from Lys-354 to Asn-408, 21 out of 55 residues (38% similarity). Overall, similarity from Ala-170 to Lys-435 is 32% (85 out of 266 residues are identical). In spite of these similarities, the M. xanthus ORF does not have the domain, which shows apparent sequence similarity with ribonuclease H (RNase H). The RNase H domain is found to be located in the carboxyl terminal region of the same polypeptide in which the RT domain exists in the amino terminal region in the case of the E. coli RT and other retroviral RTs. In the preceding paper, it was shown that there is a precise coupling between RT and RNase H activity (Lampson et al., 1989a). Therefore, RNase H may still reside with the ORF, or RNase H may be encoded by a separate gene.

Sequence Similarity with Other Proteins

In contrast to the E. coli RT and other retroviral RTs, the ORF found in M. xanthus has a long amino terminal extra domain consisting of approximately 170 residues. Interestingly, this region shows some sequence similarities with the carboxyl terminal region associated with integration protein of Mo-MLV (Moloney murine leukemia virus; Shinnick et al., 1981) (see Figure 4A); the sequence from Pro-18 to Leu-128 of the ORF shows 22% similarity (24 out of 111 residues) with the region from Pro-1070 to Leu-1179 of the gag-pol polyprotein of Mo-MLV. It should be noted that this region of Mo-MLV is unique for Mo-MLV integration protein and does not share sequence similarity with other retroviral endonucleases (Johnson et al., 1986). It is also interesting to notice that in Ty retrotransposon, this domain is located in front of the RT domain in contrast to the retroviral endonuclease domain (Clare and Farabaugh, 1985).

As pointed out above, the ORF does not have homology to E. coli or retroviral RNase H. Instead, it has a short sequence of approximately 80 residues after the RT domain. In this region, one can also find sequence similarity with a part of the gag region of HIV. As shown in Figure 4B, the sequence from Gly-411 to Glu-485 has 22 identical amino acid residues (31% similarity) with the region from Gly-396 to Pro-461 of the gag protein of HIV (Ratner et al., 1985).

Requirement of Reverse Transcriptase

The fact that disruption of the ORF significantly reduced msDNA production in *M. xanthus* (Dhundale *et al.*, 1988a) and the fact that the ORF has sequence similarity with retroviral RTs strongly supports the previous hypothesis that RT is required for the synthesis of msDNA (Dhundale *et al.*, 1987). Recently, we were able to demonstrate that msDNA is indeed synthesized by reverse transcriptase activity in a cell-free system (Lampson *et al.*, 1989a). The fact that a small amount of msDNA (3% of the wild type level) is still produced in the ORF mutants (Dhundale *et al.*, 1988a) is most likely due to another RT associated with smaller msDNA (msDNA-Mx65; previously assigned mrDNA; Dhundale *et al.*, 1988b). In fact, an ORF has been found to be associated with the region responsible for msDNA-Mx65 production.

At present it is unknown if the ORF is transcribed together with msdRNA from a common upstream promoter or if the ORF has its own independent promoter. Previously, a major RNA transcript of approximately 375 bases by S1 mapping (Dhundale *et al.*, 1987) was identified. This transcript covers the region from approximately 75 bases upstream of *msr* (at around residue-256 in Figure 2) to approximately 70 bases upstream of *msd* (at around residue-632 in Figure 2). This indicates that this RNA transcript ends at the ribosome binding site (AGG, 630-632) of the ORF. It is possible that the primary RNA transcript covers not only the *msr-msd* region but also the entire ORF. This transcript of approximately at least 2 kilobases (kb) is then used as the mRNA for the ORF to produce RT. At the same time, the 5' untranslated region of 350 bases forms a stable secondary structure which serves as a primer and a template for msDNA synthesis as previously proposed (Dhundale *et al.*, 1987). Because of the secondary structure, the 5' end region is probably much more stable than the ORF mRNA region. As a result, only the 375-base RNA from the 5' end of the transcript was detected in the previous work. In *E. coli*, the RT gene was shown to be transcribed from a single promoter for the *msr* region (Lampson *et al.*, 1989b).

Evolution of Reverse Transcriptase

All of the RTs so far identified are from eukaryotic origins, and associated with either retroviruses or retrotransponsons. DNA synthesis for retroviruses and transposition events for retrotransponsons occur via RNA which is used as a template for RTs (see review by Varmus, 1985). From amino acid similarity in various RTs, possible evolutionary relationships among these RTs has been proposed (Yuki et al., 1986).

The present invention demonstrates that RTs are not specific to eukaryotes but exist in prokaryotes as well. An intriguing question arises as to the evolutionary relationship between prokaryotic and eukaryotic RTs and the origin of RT. In order to compare the amino acid sequences of these RTs, the sequence of the M. xanthus RT from Gly-304 to Leu-371 was chosen, since this sequence includes the YXDD box, the most conserved region among different RTs. In Figure 5A this sequence is compared with 13 other representative RTs from bacteria, yeast, plant, mitochondrial plasmid, and animal retroviruses. Within these 14 sequences, the D-D sequence (residues-346 and -347) is completely conserved, and both G-311 and Y-344 are also well conserved except for Ty-RT. Besides these residues, L-308, P-309, Q-310, S-315, P-316, L-330, S-351, and L-371 are fairly well conserved among these sequences. On the basis of the numbers of identical amino acid residues, M. xanthus RT has the following similarities with other RTs: 47% (32 amino acid residues) with E. coli C1-1 RT; 41% (28) with E. coli B RT; 24% (16) with HIV, BLV, and mitochondrial plasmid RTs; 22% (15) with Mo-MLV RT; 21% (14) with RSV, 17.6, gypsy, and Tal-3 RTs; 19% (13) with HTLV1 RT; 15% (10) with Ty912 RT; and 9% (6) with Copia RT. On the basis of the phylogenetic relationships among RTs proposed by Yuki et al. (1986), and the present data, a dendrogram of homology of various RTs may be constructed as shown in Figure 5B. As proposed earlier (Yuki et al., 1986), modern RTs are composed to two major groups I and II. One group (group II) consists of retrotransponsons found in yeast (Ty912), plant (Tal-3), and Drosophila (Copia). Bacterial RTs seem to belong to the other group (group I) together with other retrotransponsons from Drosophila such as 17.6 and gypsy, mitochondrial plasmid RT, and retroviral RTs. This indicates that both prokaryotic and eukaryotic RT genes were possibly derived from a single ancestral RT gene.

Origin of the M. xanthus Reverse Transcriptase

In addition to the sequence similarity between the M. xanthus RT and RTs from retroviruses and retrotransponsons, msDNA shares other interesting similarities with retroviruses and retrotransponsons; msDNA (synthesis of single-stranded DNA) starts at a site 77 bases upstream of the RT gene and the orientation of DNA synthesis is opposite to the direction of translation of the RT gene. In the case of retroviruses and retrotransponsons, single-stranded DNA synthesis proceeds at the 5'-end untranslated region of an RNA molecule which serves as the mRNA for RT as well (Weiss et al., 1985). The orientation of DNA synthesis is also opposite to the direction of translation of the RT gene. In the case of msDNA synthesis an RNA transcript itself serving as a template also serves as a primer by self-annealing to form a stable secondary structure (Dhundale et al., 1987), whereas in the case of retroviruses and retrotransponsons tRNAs are recruited from the cell for the priming reaction. At present it is unknown if branched RNA-linked msDNA is the final product of an unknown function or if it is a stable intermediate leading to other products.

Furthermore, it is of great interest whether the M. xanthus RT is associated with a complex such as virus-like particles such as those found for yeast Ty1 element (Eichinger and Boeke, 1988). In a preliminary experiment, msDNA of M. xanthus exists as a complex with proteins in the cell which sediments as a 22S particle. Characterization of this complex may shed light on questions concerning the relationship between msDNA and retrocomponents as well as the functions of msDNA.

At present, there is no information to support the possibility that msDNA may be a transposable element or an element associated with a provirus (or prophages). It is important to point out that the RT gene from M. xanthus appears to be as old as other genomic genes for the following reasons: (a) Nine independent natural isolates of M. xanthus from various sites (including Fiji Island and eight different sites in the United States) contained mutually hybridizable msDNA (Dhundale et al., 1985). Since under the same hybridization condition, msDNA-Mx162 did not hybridize with msDNA-Sa163 [which has extensive homology in both DNA and RNA sequences with msDNA-Mx162; Dhundale et al., (1987)], the nine independent strains M. xanthus are assumed to contain almost identical msDNA. (b) The codon usage of the Mx-162 RT is almost identical to those found in other M. xanthus genes (Table 1). M. xanthus is known to have a very high G+C content (70%; Johnson and Ordal, 1968) and as a result, all the genes so far characterized have very high G+C contents at the third positions of codons used; 85.4% for vegA (Komano et al., 1987), 85.7% of ops (Inouye et al., 1983), 87.2% for tps (Inouye et al, 1983), 88.4% for mbhA (Romeo et al., 1986), and 93.9% for sigma factor. The average G+C content of the third positions is calculated to be 90.0% for these genes (Table 1). Surprisingly, the G+C content of the third positions of the RT codons is highest among these genes (95.5%; Table 1).

In contrast, the E. coli msDNA system including the RT gene is considered to have been acquired much later in the evolution of E. coli. Reasons for this conclusion include: (a) Only four strains out of 89 independent clinical E. coli strains were found to produce msDNAs (Lampson et al., 1989b). (b) The codon usage of the E. coli RT is significantly different from the general codon usage of E. coli genes obtained from 199 E. coli genes (Maruyama et al., 1986). In particular, out of 62 arginine codons used in the E. coli RT, 40 (65%) use AGA or AGG in contrast to 2.7% for the AGA+AGG usage among all arginine codons in 199 E. coli genes (see Table 1). The AGA and AGG codons are the least used codons in E. coli (Maruyama et al., 1986). In addition to AGA and AGG codons, many other codons, GCC and GCG for Ala, CGU and CGC for Arg, CAG for Gln, GGC and GGA for Gly, CAC for His, AUC and AUA for Ile, UUA, CUU and CUG for Leu, UUC for Phe, CCU and CCG for Pro, UCG for Ser, ACC and ACA for Thr, and GUC for Val. (c) Although the E. coli msDNAs share little sequence homology, they all share the key secondary structures of a branched rG residue, a DNA-RNA hybrid at the 3' ends of the msDNA and msdRNA, and stem-and-loop structures in RNA and DNA strands (Lampson et al., 1989b; Lim and Maas, 1989).

These results clearly demonstrate distinct differences between the msDNA systems of E. coli and M. xanthus. Myxobacteria are common organisms in soil and are found all over the world regardless of climate, and considered to diverge from their nearest bacterial relatives about $2x10^9$ years ago when the atmosphere became aerobic (see a review by Kaiser, 1986). Since it is reasonable to assume that the M. xanthus RT gene is as old as other genomic genes, the RT gene existed much before eukaryotic cells appeared (1.5-0.9 x 10^9 years ago). The relatedness between various prokaryotic and eukaryotic RTs as shown in Figures 5A and B strongly supports the existence of a single ancestral gene for all RTs. It is possible that such an ancestral RT gene was independently recruited into different systems such as the msDNA system, the retrotransposon system, and the retroviral system. Alternatively, the msDNA system may be a primitive ancestral system from which retrotransposons and retroviruses originated. In this regard, it is intriguing to point out other sequence similarities between the M. xanthus RT-ORF and other retroelements (see Figure 4) other than RT itself as well as the similar mode of initiation of DNA synthesis by RT as discussed earlier.

At present, it is beyond our speculation why the E. coli msDNA systems are so diverged in contrast to the M. xanthus msDNA system and how they were acquired into the genomes of some E. coli strains. However, it should be noted that the E. coli RTs are most related to the M. xanthus RT indicating that they were not derived from eukaryotic origins. Possible origins of retroviruses have been discussed (Temin, 1980). The recent finding of an imposon in a genetic component for a mouse gene also raises an interesting question concerning the evolution of retroelements (Stavenhagen and Robins, 1988). Further characterization of the prokaryotic RTs and the msDNA systems will provide clues to the origins of RT and other retroelements.

EXPERIMENTAL PROCEDURE

DNA Manipulation and Plasmids

DNA manipulation was performed as described by Maniatis et al. (1982). The plasmid isolation was as originally described by Birnboim and Dolly (1979). Plasmid pmsSB7 containing the 5 kb SalI-BamHI fragment shown between the SalI and BamHI sites of pUC9 (Vieira and Messing, 1982) was used. After the 2.2 kb SalI-SmaI fragment from pmsSB7 was subcloned between the SalI and SmaI sites of pUC9, all RsaI fragments were gel-purified and cloned into pUC9 for DNA sequence.

DNA sequence

DNA sequence was determined by the chain termination method (Sanger et al., 1977) using single-stranded or double-stranded DNA as templates with synthetic oligonucleotides.

Other Material and Methods

Restriction enzymes were purchased from either Bethesda Research Laboratories or New England BioLabs. [α-^{35}S] dATP was from Amersham. Sequenase, Version 2.0 Kit was purchased from United States Biochemical Corporation for DNA sequences.

Cyborg program from International Biotechnologies Inc. was used to search sequence homology in GenBank Release 55.

Screening of bacteria for retron synthesized msDNAs was performed by the methods of Lampson et al. J. Bacteriol, 173:5363-5370 (1991), or Yee et al, Cell, 38, 203-209 (1984).

RTs were identified and isolated by the method of Lampson et al, J. Biol. Chem, 265:8490-8496.

msDNA in Escherichia coli

The recent serendipitous finding of msDNA (msDNA-Ec86) in E. coli B by Dongbin Lim and Werner Maas (D. Lim et al., 1989) prompted a to search for msDNA in other E. coli strains. Previously established by Yee et al. (T. Yee et al., 1984), msDNA is not found in the common laboratory strain K12, however, to our surprise, it was in a clinical E. coli strain isolated from a patient with a urinary tract infection. Fifty independent E. coli urinary tract isolates were examined for the presence of msDNA (The clinical E. coli strains were urinary tract isolates kindly provided by Dr. Melvin Weinstein from the microbiology laboratory, R.W. Johnson Hospital, New Brunswick, NJ. The clinical strain Cl-1 was identified using the API-20E identification system (API laboratory products) and gave a typical E. coli profile number of 5044552.). The screening method involved treatment of total RNA prepared from each strain with (AMV) RT in the presence of [α-^{32}P]dCTP plus dATP, dTTP, and dGTP followed by polyacrylamide gel electrophoresis. Since msDNA contains a DNA-RNA duplex structure, the 3' end of the DNA molecule serves as an intramolecular primer and the RNA molecule as a template for RT. When RNA prepared from one of the clinical strains, E. coli Cl-1, was labeled in this manner, two distinct, low molecular weight bands of about 160 bases became labeled with ^{32}P and are shown in Figure 6. If the labeled sample is digested with ribonuclease (RNase) A prior to loading on the gel, a single band corresponding to 105 bases of single-stranded DNA is detected (lane 4). This indicates that both bands in lane 3 contain a single-stranded DNA of identical size. The two labeled bands observed prior to RNase treatment (lane 3) are due to two species of msDNA comprised of a single species of single-stranded DNA linked to RNA molecules of two different sizes. RNA molecules of two different sizes have been observed at the 5' ends of msDNA from myxobacteria in which a precursor molecule contains a longer RNA which is processed into a smaller mature form (Dhundale et al., 1987; Furuichi et al., 1987). Among the 89 clinical isolates screened, three other strains produced msDNA-like molecules of varying size and quantity, suggesting extensive diversity among these molecules. As previously reported (Dhundale, 1985), msDNA was not observed in the E. coli K-12 strain, C600 (lanes 1 and 2, Figure 6).

Nucleotide sequence of msDNA Ec-67

To determine the base sequence of the DNA molecule, the RNA-DNA complex isolated from the clinical strain was labeled at the 3' end of the DNA molecule with AMV-RT and [α-^{32}P]dATP. By adding dideoxy-CTP, ddTTP, and ddGTP to the reaction mixture, a single labeled adenine is added to the 3' end of the DNA molecule. RNA is removed with RNase A+ T1 and the end-labeled DNA is subjected to the Maxam and Gilbert sequencing method (Maxam et al., 1980). Figure 7 shows that msDNA consists of a single-stranded DNA of 67 bases and, as in the case of msDNAs from myxobacteria (Yee, 1984; Dhundale, 1987), it can form a secondary hair-pin structure. The primary sequence, however, is not homologous to any of the myxobacterial msDNAs, nor to the msDNA from E. coli B (msDNA-Ec86; Lim and Maas, personal communication).

The sequence of the RNA molecule was determined using the RNA-DNA complex purified from E. coli Cl-1. The RNA sequence was determined using base specific RNases as described previously (Dhundale et al., 1988). As shown in Figure 8, a large gap is observed in the RNA sequence "ladder". This gap is due to the DNA strand branched at the 2' position of the 15th rG residue of the RNA strand which produces a shift in mobility of the sequence ladder (see Figure 7). The RNA consists of 58 bases with the DNA molecule branched at the G residue at position 15 by a 2',5'-phosphodiester linkage. The branched G structure was determined as previously described for msDNAs from myxobacteria (Dhundale, 1987; Furuichi et al., 1987). After RNase (A and T1) treatment, msDNA retains a small oligoribonucleotide linked to the 5' end of the DNA molecule due to the inability of RNases to cleave in the vicinity of the branched linkage. The 5' end was labeled with [Υ-^{32}P]ATP using T4 polynucleotide kinase and the labeled RNA molecule was detached from the DNA strand by a debranching enzyme purified from HeLa cells (Ruskin et al. 1985; Arenas et al., 1987; the debranching enzyme was a gift from Jerard Hurwitz). This small RNA was found to be a tetraribonucleotide which could be digested with RNase T1 to yield a labeled dinucleotide (not shown). Since RNase T1 could not cleave the RNA molecule at the G residue before debranching enzyme treatment, it was concluded that the single-stranded DNA is branched at the G residue via a 2',5'-phosphodiester linkage. In addition, partial RNase U_2 digestion cleaved the RNA molecule to yield a ^{32}P-labeled mono- and a ^{32}P-labeled trinucleotide (not shown). Thus, the sequence of the tetranucleotide is $^{5'}$A-G-A-(U or C)$^{3'}$. Based on these data, the complete structure of msDNA-Ec67 from E. coli Cl-1 is presented in Figure 7. Despite a lack of primary structural homology, msDNA-Ec67 displays all the unique features found in msDNAs from myxobacteria. These include a single-stranded DNA with a stem-and-loop structure, a single-stranded RNA with a stem-and-loop structure, a 2',5'-phosphodiester linkage between the RNA and DNA, and a DNA-RNA hybrid at their 3' ends. This hybrid structure was confirmed by demonstrating sensitivity of the RNA molecule to RNase H (not shown).

Cloning of the locus for msDNA-Ec67

In order to identify the DNA fragment which is responsible for msDNA synthesis in E. coli Cl-1, Southern blot hybridization was carried out with various restriction enzyme digests of total chromosomal DNA prepared from E. coli Cl-1, using msDNA-Ec67 labeled with AMV-RT (the same preparation as shown in lane 3, Figure 6) as a probe. The result is shown in Figure 9A. EcoRI (lane 1), HindIII (lane 2), BamHI (lane 3), PstI (lane 4) and Bg11II (lane 5) digestions showed single band hybridization signals corresponding to 11.6, 2.0, ˷22, 2.8 and 2.5 kilobase pairs (kb), respectively. The upper band appearing in the EcoRI digestion is due to incomplete digestion of the chromosomal DNA. Analysis of total chromosomal DNA prepared from E. coli Cl-1 by agarose gel electrophoresis revealed that the strain contains two plasmids of different size. However, neither plasmid hybridized with the ^{32}P- labeled probe, indicating that the fragments detected in Figure 9A are derived from chromosomal DNA. Furthermore, there is only one location for the msDNA-coding region on the chromosome, since various restriction enzyme digestions gave only one band of varying sizes. Similar results were observed for the msDNAs of myxobacteria (Yee et al., 1984; Furuichi et al., 1987; and Dhundale et al., 1988).

The 11.6-kb EcoRI fragment and the 2.8-kb PstI fragment were each cloned into pUC9 (Yanisch-Perron et al., 1985) and E. coli CL83 (a recA transductant of strain JM83), an msDNA-free K-12 strain (lane 1, Figure 9B), was transformed with the plasmids. Cells transformed with the 11.6-kb EcoRI clone (pCl-1E) were found to produce msDNA (lane 2, Figure 9B), whereas cells transformed with the 2.8-kb PstI clone (pCl-1P) failed to produce any detectable msDNA (lane 3, Figure 9B). A map of the 11.6-kb fragment is shown in Figure 10. Southern blot analysis of the fragment revealed that a 1.8-kb PstI - HindIII fragment hybridized with the msDNA probe. When the DNA sequence of this fragment was determined, a region identical to the sequence of the msDNA molecule was discovered. The DNA sequence corresponding to the sequence of msDNA is indicated by the enclosed box on the lower strand in Figure 11 and the orientation is from right to left. The location of this sequence is also indicated by a small arrow in Figure 10. As is the case for all other known myxobacterial msDNAs (Dhundale et al., 1987; Furuichi et al., 1987; and Dhundale et al., 1988), a sequence identical to that of the RNA linked to msDNA (see Figure 7) was found downstream of the msDNA-coding region in opposite orientation and overlapping with that region by 7 bases. This sequence is indicated by the enclosed box on the upper strand in Figure 11 and the branched G residue is circled. Again, as in all the msDNAs found in myxobacteria, there is an inverted repeat comprised of a 13-base sequence immediately upstream of the branched G residue (residue 250 to 262; sequence a2 in Figure 11) and a sequence at the 3' end shown by an arrow in Figure 11 (residue 368 to 380; sequence a1). As a result of this inverted repeat, a putative longer primary RNA transcript beginning upstream of the RNA coding region and extending through the msDNA coding region would be able to self-anneal and form a stable secondary structure, which is proposed to serve as the primer as well as the template for msDNA synthesis (Dhundale et al., 1987).

Existence of an essential gene for msDNA synthesis

The 2.8-kb PstI fragment (from PstI(a) to PstI(b) in Figure 10) was not able to synthesize msDNA. However, an overlapping 3.9-kb fragment from BalI (1.0 kb downstream of PstI(a); see Figure 10) to the following EcoRI site contains all the information required for synthesis of msDNA. This indicates that a region downstream of the PstI(b) site (Figure 10) is required for msDNA production. The nucleotide base sequence from this region revealed a long open reading frame (ORF) of 586 amino acid residues, starting with the initiation codon ATG at nucleotide 418 to 420 as shown in Figure 11. A distance of only 51 bases separates the initiation codon from the region which encodes msDNA. A putative Shine-Dalgarno sequence (GGA) can be found 10 bases upstream of the initiation codon. When the lacZ gene was fused in frame at the HindIII site (within the ORF) at amino acid residue-126, β-galactosidase activity was detected (not shown). Thus the region encompassing the ORF is indeed transcribed and the gene product encoded by the ORF is essential for msDNA synthesis. In a preliminary experiment, both msdRNA and the ORF appeared to be transcribed as the same transcription unit, since a deletion mutation removing the sequence from residue 1 to 181 blocked the expression of the lacZ gene fused at the HindIII site. A putative promoter can be found in the deleted sequence as boxed in Figure 11. These -35 and -10 regions probably serve as the promoter for both msdRNA synthesis and the ORF.

Sequence similarity with retroviral reverse transcriptases

When the amino acid sequence of the ORF was compared with known proteins, a striking similarity was found with retroviral RTs. In Figure 12, the ORF is compared with RTs from HIV (human immunodeficiency virus; Ratner et al., 1985; and Johnson et al., 1986), and HTLV1 (human T-cell leukemia virus type I; Seiki et al., 1983; and Patarca et al., 1984). The first domain (Asn-32 to Val-291) matches well with the RT domains of HIV and HTLV1. In particular, the sequences around the polymerase consensus "Asp-Asp" sequence (Toh et al., 1983; and Geng et al., 1985; boxed in Figures 11 and 12) are well conserved. Out of 260 amino acid residues in this domain, 44 and 38 residues are identical with HIV and HTLV1, respectively. Between HIV-RT and HTLV1-RT, there are 78 identical amino acid residues in this domain.

The pol gene of retroviruses is known to produce a protein consisting of RT and RNase H activities; the former at the amino-terminal and the latter at the carboxyl-terminal region of the pol gene product (Ratner et al., 1985; Johnson et al., 1986; Varmus, 1985; and Tanese et al., 1988). These domains have been shown to be separated by a poorly conserved "tether" domain of approximately 160 to 190 amino acid residues (Ratner et al., 1985; Johnson et al., 1986). On the basis of the HIV sequence, the similarities (only identical amino acid residues) between HIV and HTLV1 are 29.5 and 16.8% for the RT domain and the tether domain, respectively. The similarities between HIV and msDNA are 16.9 and 10.3% for the RT domain and the tether domain, respectively. The similarities between HTLV1 and msDNA are 14.6 and 15.5% for the RT domain and the tether domain, respectively. These results indicate that in addition to the RT region, there are reasonable similarities in the tether domain between retroviruses and msDNA. An alignment of the RNase H domains also revealed that there are similarities between retroviruses and msDNA (15.7 and 17.4% with HIV and HTLV, respectively; see Figure 12). The similarity between HIV and HTLV1 in this region is 18.0%.

Cell extracts were prepared and assayed for the presence of RT activity associated with the production of msDNA as predicted from the amino acid homologies. Only the E. coli strain (C2110, polA) (Tanese et al., 1985; Tanese et al., 1986; E. coli strain C2110 (polA1⁻) was a gift from M. Roth and S. Goff) harboring the plasmid, pC1-1EP5, containing the msDNA ORF displayed RT activity (Figure 13). The polA strain was used to eliminate high background activity in the RT assay due to DNA polymerase I. No RT activity was detected in extracts containing the vector plasmid alone, or when the template-primer (poly rC-dG) was absent from the reaction mix (Figure 13). It is interesting to note that the PstI(b) site is located at amino acid residue-430, which is between the tether domain and the RNase H domain. A plasmid lacking sequences downstream of the PstI(b) site did not produce msDNA. This suggests that the RNase H domain may be essential for msDNA synthesis, or alternatively that PstI disruption may result in inactivation of RT.

In addition to the similarity between msDNA-Ec67 RT and retroviral RT, there is an interesting similarity between msDNA and retroviruses; DNA synthesis starts at a site upstream of the RT-RNase H gene, and the orientation of DNA synthesis is opposite to the direction of transcription of the RT-RNase H gene. In the case of retroviruses, tRNAs are recruited from the cell for the priming reaction (Weiss et al., 1985), whereas for msDNA an RNA transcript serving as, template also serves as a primer by self-annealing to form a stable secondary structure (Dhundale et al., 1987; Furuichi et al., 1987).

Origin of the E. coli Reverse Transcriptase

At present the relationship between msDNA and retroviruses is an open question. It is possible that the study of msDNA may shed light on the question of the origin and evolution of retroviruses. It is an intriguing question to consider why some of the clinical E. coli strains, isolated from human patients produce msDNA. Our preliminary data indicate that msDNAs produced by four independent E. coli strains, isolated from urinary track infections, share little homology. This suggests that there may be enormously large numbers of species of msDNA in E. coli. In contrast to msDNAs found in E. coli, msDNA-Mx162 from M. xanthus is highly conserved, since nine independent M. xanthus strains isolated from various sites have msDNA which hybridizes with the original msDNA-Mx162 (Dhundale et al., 1985). Furthermore, msDNA from another myxobacterium, S. aurantiaca (msDNA-Sa163; Furuichi et al., 1987), also shows a high degree of homology to msDNA-Mx162 (Furuichi et al., 1987).

Several lines of evidence suggest that the RT gene found in the E. coli strain Cl-1 is not likely to have originated in E. coli, but rather was recently acquired from some other source. For example, only about 4% of E. coli strains tested were found to produce msDNA. In addition, the RT gene from strain Cl-1 does not cross hybridize to chromosomal DNA from four other E. coli strains which produce msDNA molecules, indicating that there is extensive diversity among these RT genes. In contrast, a DNA fragment from the E. coli-K-12 sigma factor gene can hybridize to chromosomal DNA from all five msDNA producing, E. coli strains, indicating the conserved nature of sigma factors. An analysis of the E. coli RT gene indicates that the codon usage for this gene is remarkably different from most E. coli proteins. In particular, AGA and AGG, the least frequently (2.7%) used codons for arginine among 199 E. coli genes (Maruyama et al., 1986), occurs at a frequency of 64.5% in the E. coli RT gene. Similarly, CUG is the most commonly used codon for leucine (61.3%; Maruyama et al., 1986) in E. coli genes, while its prevalence in the RT gene is only 9.1%. The AT base pair content of the E. coli RT gene was calculated to be 67.6%, which is substantially higher than the AT content of the E. coli genome (45%; Fasman, 1976). The AT contents of HIV and HTLV1 RT genes are 62.1% and 47.8%, respectively. These facts pose an intriguing question as to how and when the RT gene, as well as the msDNA coding region, were integrated into the genome of the clinical strain.

There are many questions to be answered, including (a) are there any particles associated with msDNA, (b) is the msDNA region transposable like the Ty element of yeast (Boeke et al., 1985; Eichinger et al., 1988), (c) can the element responsible for the production of msDNA be transferred from cell to cell, (d) can a RT from one strain (E. coli or myxobacteria) complement the production of msDNA of other strains, (e) does the promoter for the RNA transcript have any similarities to the retroviral LTR, (f) are there any specific integration sites for the msDNA element on the E. coli chromosome, (g) why is the branched G residue conserved, (h) is there an enzyme responsible for priming DNA synthesis at the 2'-OH position of the rG residue, (i) why and how does msDNA synthesis stop at one distinct site on the RNA template, and (j) how different biochemically are the msDNA RTs from retroviral RTs?

The existence of reverse transcriptase in prokaryotes, previously speculated upon (Dhundale et al., 1987), is now evident. This fact raises intriguing questions concerning possible roles of this enzyme in the prokaryotes other than a role in msDNA production. Recently we also found that M. xanthus, in which msDNA was originally discovered, has a long ORF in the same manner as found for msDNA-Ec67. This ORF has a high degree of similarity to the E. coli RT. Since eight independent isolates of M. xanthus produce homologous msDNA, the M. xanthus RT is likely to have been acquired at a very early stage of its evolution in contrast to the E. coli RT. The determination of the structures of both M. xanthus and other E. coli RTs will shed light on the key question of the origin of RT and its role in prokaryotes.

An important embodiment of the invention relates to the discovery of msDNA-producing retron elements in a number of diverse bacterial groups. Thus, retron elements appear to be widely prevalent, at least amongst the purple bacteria or proteobacteria including Proteus, Klebsiella and Salmonella of the gamma subdivision; Rhizobium and Bradyrhizobium from the alpha subdivision; and Nannocystis (a myxobacterium) from the delta subdivisions. These are representatives of the three of the four major subdivisions of the purple bacteria of proteobacteria. As shown above the retron-encoded RT is responsible for the synthesis of msDNAs.

The retron elements were discovered by detecting the presence of msDNA by one of two classic methods: the so-called "RT extension method", described by Lampson, B.C., M. Inouye and S. Inouye, 1991. Survey of multicopy single-stranded DNAs and reverse transcriptase genes among natural isolates of Myxococcus xanthus. J. Bacteriol. 173:5363-5370 and in Lampson, B.C., M. Viswanathan, M. Inouye and S. Inouye, 1990. Reverse transcriptase from Escherichia coli exists as a complex with msDNA and is able to synthesize double-stranded DNA. J. Biol. Chem. 265:8490-8496 or polyacrylamide gel electrophoresis of a chromosomal DNA extract followed by staining with ethidium bromide as described by Yee, T., T. Furuichi, S. Inouye, 1984. Multicopy Single-Stranded DNA Isolated from a Gram-Negative Bacterium, Myxococcus xanthus. Cell, Vol. 38, 203-209. Both of these publications are incorporated herein by reference. Both methods provide a reliable, convenient and conventional protocol for screening of bacteria for the presence of retron-encoded RT and msDNAs.

In accordance with the RT extension method, the DNA portion of msDNA is specifically ^{32}P radio labeled. Radio labeled from a total RNA preparation extracted from each bacteria strain to be screened. Twenty or more isolates of proteus mirabilia, Klebsiella pneumoniae, Salmonella species, rhizobial species, and enterococcal species were screened by this method. Low-molecular-weight bands (Fig. 20) indicated the presence of small labeled DNAs after polyacrylamide gel electrophoresis and autoradiography of the labeling reaction mixes. In addition, half of each labeling reaction mix was also treated with RNase A, causing a shift to a faster-migrating band, indicating that the labeled DNA is also associated with RNA. This is hallmark of the msDNA molecule as discussed above. Four of the 23 P. mirabilia isolates screened produced msDNA, while only 1 of 21 K. pneumoniae isolates and 4 of 70 Salmonella isolates screened produced msDNA. msDNA was detected in any of the 30 or so enterococcal strains screened by this method. It was concluded that the bacterial genera which contain msDNA producing retron elements are representatives of three of the four major subdivision of the purple bacteria or Proteobacteria, as described above.

In accordance with this embodiment of the invention, it is noteworthy that the discovery of msDNA extends for the first time the distribution of retron-elements to a new phylogenetic division of the purple bacteria, namely, the alpha subdivision. A collection of 63 rhizobial isolates (shown in Table 1) were screened for the presence of msDNA by the RT extension method. Among the 63 isolates, msDNA were detected in 10 (16% - Fig. 20 and Fig. 21). However, all 10 positive isolates give strong, clearly labeled bands with a typical shaft of a fast-migrating band after treatment with RNase A, indicating the presence of RNA and DNA in the labeled molecule. The 10 retron-encoding rhizobial strains include both fast growing (rhizobium) and slow-growing

(Bradyrhizobium) rhizobia.

The RT extension method comprises treating a preparation of total RNA, extracted from a bacterial strain to be tested, with RT from a suitable source in the presence of the deoxynucleotides dATP, dTTP, dGTP and dCTP, one of which is radiolabeled, e.g., $[\alpha\text{-}^{32}P]$ dCTP, electrophoresing the treated RNA preparation on a polyacrylamide gel and determining initially the presence or absence of msDNA in the bacterium of interest by detecting a band of radiolabeled DNA corresponding to the single-stranded DNA of msDNA. Typical examples of suitable sources of RT are avian myeloblastosis virus (AMV) and Moloney murine leukemia virus (Mo-MLV). Conceivably, the test could be automated.

Total RNA samples, which contain msDNA if present in the bacterium, are extracted from the bacterial strain of interest and prepared for RT extension as follows. Total RNA, prepared from a 5-ml culture from the bacterial strain, is added to 50 μl of a reaction mixture containing: 50 mM tris-HCl (pH 8.3); 6 mM $MgCl_2$; 40 mM KCl; 5 mM DTT; 1 μm dATP, dTTP and dGTP; 0.04 μM dCTP; 0.2 μM $[\alpha\ ^{32}P]$ dCTP; and 10 units of AMV-RT (Boehringer Mannheim). The reaction mixture is incubated at 37^0C for 30 minutes, then extracted with 50 μl of phenolchloroform (1:1) and precipitated with ethanol. The samples are subjected to electrophoresis on a 4% acrylamide -8 M urea gel with appropriate nucleotide size markers, e.g., the Klenow fragment of DNA polymerase I. If the labeled sample is digested with ribonuclease (RNase) A before it is placed on the gel, a single band corresponding to single-stranded DNA is detected, which is indicative of the presence of msDNA. An aliquot from each labeling reaction mixture is treated with 5 μg of RNase for 10 minutes at 37^0C just prior to electrophoresis to detect in the gel a shift to a faster - migrating species, indicating that each labeled DNA is also associated with RNA, which is the hallmark of the msDNA molecule.

Low-molecular weight bands in the gel indicate the presence of small labeled DNAs after polyacrylamide gel electrophoresis and autoradiography of the labeling reaction mixtures.

Multiple bands observed in some of the lanes of the gel even after RNase treatment may be due to incomplete extension by RT during the labeling reaction, or, alternatively, multiple forms or species of msDNA may exist in a given bacterium.

The Yee method for screening bacteria for the presence of retrons which synthesize msDNAs involves purifying by a conventional phenol extraction procedure total chromosomal DNA from the desired bacteria to be screened, electrophoresis on a five percent preparation acrymalide gel and checking for a satellite band. The major satellite band is cut out to extract the material in the band to quantitate the material in the satellite band. Total chromosomal DNA is subjected to acrylamide gel electrophoresis, the gel is stained with a ethidium bromide and densitometric scanning is employed to quantitate the satellite DNA against the pBR322 standard. The method is described in better details in Yee cited above.

A collection of rhizobial isolates from the United States Department of Agriculture (USDA) Beltsville Rhizobium Culture Collection are screened for the presence of msDNA by the RT extension method. This collection represents isolates at different times, from different legume hosts and from different geographic locations. msDNAs are detected in 10 isolates. All 10 positive isolates give strong, clearly labeled bands of DNA, with a typical shift to a fast-migrating band after treatment with RNase A, indicating the presence of RNA and DNA in the labeled molecule. The 10 retron-encoding rhizobial strains include both fast-growing (Rhizobium) and slow-growing (Bradyrhizobium) rhizobia as follows: Rhizobium sp. (Acacia) 3002 and 3838, Bradyrhizobium sp. (Aeschynomene) 3516, Bradyrhizobium sp. (Albizia) 3004, Bradyrhizobium sp. (Erythrima) 3242, Rhizobium loti 3468 and 3503, Rhizobium trifolii 2048 and 2065 and Bradyrhizobium sp. (Vigna) 3447. See Figure 21

Total DNA from each of eight msDNA-producing strains clearly cross-hybridizes with a nod YAB (1.6 - kb Eco RI fragment) gene probe derived from Bradyrhizobium japonicum, confirming that these strains are members of the Rhizobiaceae.

In view of the diversity of retron elements in prokaryotic populations, it is not excluded that msDNA synthesizing retrons would be found in bacteria living in alkaline environments, such as in alkaline environments: Plectonema nostocorum, Flavobacterium spp. Agrobacterium spp. Bacillus spp. Ectothiorhodospira spp.; in acidic environments: Thiobacillus thermophilica and thiooxidans, Thermoplasma acidophilus, Sulfolobus acidocaldarius, Cuanidium caldarius, Bacillus acidocaldarius; in very high temperature environment (thermophilic): Sulfolobus acaidocaldarius, Caldariella acidophila, Thermus aquaticus; in very low temperature (psychrotrophic): Vibrio marinus, Pseudomonas spp., Cytophaga spp., Flavobacterium spp.; in high salt environments (halophilic): Halobacterium cutirubrum and salinarium, Halococcus morrhuae, Danaliella viridis; in high barometric pressure (like deep sea - barophilic), which are believed to inhibit the gut of ocean bottom dwelling fish. By using one of the two screening tests identified above, one skilled in the art will readily determine whether any one of these bacteria contain retrons synthesizing msDNA. This may be particularly interesting for making evolutionary comparisons between homologous RT genes present in distantly related phytogenic strains.

A representative number of amino acid sequences of representative RTs were analyzed to determine similarities and differences. The following observations were made. The amino acid sequences of these bacterial RTs are shown in Figure 14. The individual nucleotide and amino acid sequences for each of the RTs are shown in Figures 2, 11 and 15 through 19.

From a comparison of these sequences, it is noted that there are 61 conserved positions in the RT domains as indicated by solid dots at the bottom of the sequences in Figure 14. It is further noted that all bacterial RTs possess the YXDD sequence. Several other residues are conserved including the LPQS sequence that is especially common in retroviral reverse transcriptases. The RT domains are divided into seven subdomains. For each subdomain, the consensus sequences for the seven bacterial RTs can be established, as shown at the bottom of the sequences in Figure 14. There are 18 extra residues (except 26 residues for RT-Ec67) between subdomains 2 and 3, in which there is a reasonably good consensus sequence.

It has been noted that the RTs of the present invention possess a number of common conserved sequences of nucleotides and amino acid residues.

The most common conserved sequence of amino acid residues noted is as follows: tyrosine, alanine or cysteine and two aspartic acid residues. This conserved sequence, common to all
, as shown in Seq. ID No. 4350
RTs of the present invention, is also known as the YXDD sequence.

A second conserved sequence of amino acid residues noted is as follows: serine, x which is a hydrophobic residue selected from the group consisting of valine, phenylalanine leucine and isoleucine, x_1 which is a polar residue selected from the group consisting of threonine, asparagine, lysine and serine and x_2 which is a hydrophobic residue selected from the group consisting of tryptophan, phenylalanine and alanine~~.~~, as shown in Seq. ID No. ~~44~~ 51

A third conserved sequence of amino acid residues noted is as follows: asparagine, x which is a hydrophobic residue selected from the group consisting of alanine, leucine and phenylalanine and x_1 which is a hydrophobic residue selected from the group consisting of leucine, valine and isoleucine~~.~~, as shown in Seq. ID No. ~~45~~ 52

A fourth conserved sequence of amino acid residues further noted is as follows: x which is a polar residue selected from the group consisting of arginine, glutamic acid, lysine, valine and glutamine, a second residue which is valine, a third residue which is threonine and a fourth residue which is glycine~~.~~, as shown in Seq. ID No. ~~#6~~ ~~53~~ 52

These conserved sequences are only a portion of the total number of common sequences of the RTs. For other conserved sequences held in common by the bacterial RTs reference is made to Figure 14.

The RTs of the other groups of bacteria described herein as capable of synthesizing msDNAs are likewise believed to have a similar profile of conserved nucleic acid and amino acid residue sequence similarities as shown in Figure 14 and discussed above. This observation also applies to the genus Nannocystis.

In accordance with the invention, it is contemplated that prokaryotic reverse transcriptase, which is essential for msDNA synthesis, may be responsible for host cell parasitic or selfish DNA synthesis. Additionally, it is thought that the prokaryotic reverse transcriptase molecule may be essential for synthesis of biological messengers and nucleic acid enzymes.

The msDNAs synthesized by the reverse transcriptase disclosed herein possess a highly stable RNA; it is capable of self-annealing and may serve as the primer and template for msDNA synthesis. The reverse transcriptases (RTs) disclosed herein may be used as diagnostic agents. It is also contemplated that the RTs of the invention can synthesize msDNAs which will contain specific selected DNA fragments that can hybridize with complementary ssDNA, or otherwise identify ssDNAs, sought for, thus being useful as probes.

The possibility for the msDNAs to behave like restriction enzymes (or have restriction-like enzyme activity) in being capable of cleaving DNAs, or cut off a segment of itself, cannot be excluded.

The following examples are provided for purposes of illustration only and are not to be viewed as a limitation of the scope of the invention. The following examples are illustrative of bacterial isolates screened and identified to contain msDNA by way of the present invention.

EXAMPLE 1

One of the rhizobial strains, Rhizobium trifolii USDA 2065 is identified as containing msDNA by the RT extension method by which msDNA from total RNA is specifically labeled with ^{32}P as follows.

Total RNA from a 5-ml culture of R. trifolii 2065 is added to a 50 μl reaction mixture containing: 50 mM tris-HCl (pH 8.3); 6 mM Mg Cl_2; 40 mM KCl; 5 mM DTT; 1 μm dATP, dTTP and dGTP; 0.04 μMd CTP; 0.2 μM [α^{32}P] dCTP; and 10 units of AMV-RT (Boehringer Mannheim). The reaction mixture is incubated at 37^{0}C for 30 minutes, then extracted with 50 μl of phenolchloroform (1:1) and precipitated with ethanol. The samples are subjected to electrophoresis on a 4% acrylamide-8 M urea gel with appropriate nucleotide size markers, such as the Msp I digest of pBR322 end-labeled with [α-^{32}P] dCTP and the Klenow fragment of DNA polymerase I. An aliquot of the reaction mixture containing R. trifolii RNA is treated with 5 μg of RNase for 10 minutes at 37^{0}C prior to electrophoresis to detect in the gel a shift to a faster-migrating species, which indicates that the ^{32}P-labeled DNA extended by RT is also associated with RNA, which clearly demonstrates the presence of msDNA.

Low-molecular weight bands in the gel indicate the presence of small ^{32}P-labeled DNA after polyacrylamide gel electrophoresis and autoradiography. The labeled DNA is indicative of the presence of msDNA.

EXAMPLE 2

By the method described above in Example 1, (a) Proteus mirabilis 1174b is found to synthesize msDNA by the retrons containing the RT; (b) Klebsiella pneumoniae 912b is found to synthesize msDNA by RT; (c) Salmonella sp. strain SARB-3 is found to synthesize msDNA by the retrons containing the by the retrons containing the RT; (d) Nannocystis exedens Nael is found to synthesize msDNA by RT; (e) Bradyrhizobium spp. 3447, 3516 and 3004 are also found to synthesize msDNA by the retrons containing the RT.

The following method, exemplified for E. coli, for the isolation and purification of bacterial RT is applicable to bacteria which are screened as positive for the presence of msDNA by the RT extension in vitro method.

EXAMPLE 3

Isolation and Purification of Bacterial Reverse Transcriptase.

The following is a description of a convenient method for isolating and purifying a bacterial RT.

From 10 liters of a stationary phase culture of E. coli strain C2110 harboring plasmid pCl-1EP5b, cells are harvested, washed in 50 mM Tris (pH 8.0), and resuspended in lysozyme buffer (50 mM Tris (pH 7.5), 10% sucrose, 0.3 M NaCl, 1 mM EDTA, 1 mM phenylmethylsulfonyl fluoride). Fresh lysozyme is added to a final concentration of 2 mg/ml. The suspension is incubated on ice for 15 minutes followed by a quick freeze at -70^{0}C, then thawed on ice. Lysis is enhanced by the addition of 2 volumes of buffer M (50 mM Tris (pH 7.0), 1 mM dithiothreitol, 0.2% Nonidet P-40, 10% glycerol, and 25 mM NaCl) followed by incubation on ice, then a quick freeze-thaw. A cleared lysate is obtained by centrifugation at 38,000 rpm in a 50Ti rotor for 30 minutes. The cleared lysate is fractionated by ammonium sulfate precipitation (0-50%, 50-70% and 70-90%), followed by dialysis overnight (4^0C) for each fraction against buffer M. Ammonium sulfate fractions, 50-70% and 70-90%, show RT activity and are pooled, then applied to a DEAE-column (2.5 x 50 cm; DE52 Whatman) equilibrated with buffer M. The DE52 column is washed, and RT activity is eluted from the column at a range of 300 to 350 mM NaCl. The DE52 fractions showing RT activity are pooled, concentrated by membrane ultrafiltration (Amicon) and then loaded onto a Sephacryl S-300 column (Pharmacia LKB Biotechnology Inc., 1.5 x 75 cm) equilibrated with buffer M. The column is developed with the same buffer. Again, fractions from the S-300 column having RT activity are pooled and concentrated, and 0.7 ml is loaded onto a 16-30% glycerol density gradient. The glycerol gradients are set up and run as described previously (Viswanathan et al., 1989). The purified Ec67.RT (fractions 7, 8 and 9) is stored as separate glycerol fractions at -20^0C.

When this protocol is applied to the msDNA bacterial synthesizing strains, the respective RTs are isolated and identified as shown above.

Another convenient method for isolating and purifying reverse transcriptase is published in Lampson B.C., S. Inouye and M. Inouye, "msDNA of Bacteria", Progress in Nucleic Acid Research and Molecular Biology, Vol. 40, pages 1 et seq.

The invention has been described in detail with particular reference to the above embodiments. It will be understood, however, that variations and modifications can be affected within the spirit and scope of the invention.

CLAIMS

We claim:

Sub B'

1. An isolated and purified bacterial reverse transcriptase (RT) which is capable of synthesizing msDNA, which RT comprises a conserved sequence of amino acid residues as follows: tyrosine, x which is alanine or cysteine, and two aspartic acid residues.

2. The bacterial RT of claim 1 which comprises a second conserved sequence of amino acid residues as follows: serine, x which is a hydrophobic residue selected from the group consisting of valine, phenylalanine, leucine and isoleucine, x_1 which is a polar residue selected from the group consisting of threonine, asparagine, lysine and serine and x_2 which is a hydrophobic residue selected from the group consisting of tryptophan, phenylalanine and alanine.

3. The bacterial RT of claim 2 which comprises a third conserved sequence of amino acid residues as follows: asparagine, x which is a hydrophobic residue selected from the group consisting of alanine, leucine and phenylalanine and x_1 which is a hydrophobic residue selected from the group consisting of leucine, valine and isoleucine.

4. The bacterial RT of claim 3 which comprises a fourth conserved sequence of amino acid residues as follows: x which is a polar residue selected from the group consisting of arginine, glutamic acid, lysine, valine and glutamine, a second residue which is valine, a third residue which is threonine and a fourth residue which is glycine.

5. The bacterial RT of claim 1 which has the common subdomains 1 through 7 shown in Table 5.

6. The bacterial RT of claim 1 wherein the conserved sequence is located in subdomain 5 shown in Table 5.

7. The bacterial RT of claim 6 which has a total of 61 conserved amino acid residues.

8. An isolated and purified bacterial RT which comprises a sequence of amino acid residues shown in Figure 14.

9. An isolated and purified bacterial RT from a bacterium which is capable of synthesizing an msDNA as determined by the reverse transcriptase extension in vitro screening test, which indicates the presence or absence of msDNA in the bacterium.

10. The bacterial RT of claim 9 wherein the bacterium is selected from the group of genera consisting of Myxococcus, Escherichia, Proteus, Klebsiella, Flexabacter, Cytophaga, Stigmatella, Salmonella, Nannocystis, Rhizobium and Bradyrhizobium.

11. The bacterial RT of claim 10 wherein the in vitro screening test for determining the presence or absence of msDNA in the bacterium comprises treating a preparation of total RNA extracted from the bacterium with a reverse transcriptase (RT) in the presence of a radiolabeled deoxynucleotide, which RT, when msDNA is present in the total RNA of the bacterium, utilizes the DNA portion of the msDNA as a primer and the RNA portion of the msDNA as a template for radiolabeling the DNA portion of the msDNA, electrophoresing the treated RNA preparation and determining the presence of msDNA in the bacterium by detecting a band of radiolabeled DNA, said band being indicative of the presence of msDNA in the bacterium.

add B2

add E1

REFERENCES

Birnhoim H.C., and J. Dolly, Nucl. Acid Res., 7, 1513-1523 (1979).

Boeke J.D., Gorfinkel C.A., Styles C.A., Fink G.R., Cell, 40, 491 (1985).

Cairns J., Overbaugh J., Miller S., Nature, 335, 142-145 (1988).

Churchward G., Belin D., Nagaime Y., Gene, 31, 165 (1984).

Clare J., Farabaugh P., Proc. Natl. Acad. Sci. USA, 82, 2829-2833 (1985).

Dhundale A., Furuichi S., Inouye S., Inouye M., J. Bacteriol., 164, 914 (1985).

Dhundale A., Lampson B., Furuichi T., Inouye M., Inouye S., Cell, 51, 1105 (1987).

Dhundale A, Inouye M., Inouye S., J. Biol. Chem., 263, 9055 (1988).

Dhundale A., Furuichi T., Inouye M., Inouye S., J. Bacteriol., 170, 5620-5624 (1988a).

Dickson C., Eisenman R., Fan H., Hunter E., Teich N., Molecular Biology of Tumor Viruses, ed. 2, Cold Spring Harbor Laboratory NY, 513, 648 (1982).

Eichinger D.J., Boeke J.D., Cell, 54, 955-966 (1988).

Fasman G., CRC Handbook of Biochem. and Mol. Biol., Nucleic Acids, Vol 2., 102 (1976).

Furuichi T., Dhundale A., Inouye M., Inouye S., Cell, 48, 47-53 (1987a).

Furuichi T., Inouye S., Inouye M., Cell, 48, 55-62 (1987b).

Hsu M.Y., Inouye S., Inouye M., J. Biol. Chem., 264 (1989).

Inouye S., Franceschini T., Inouye M, Proc. Natl. Acad. Sci., USA, 80, 6829-6833 (1983).

Inouye, S., Hsu, M.Y., Eagle, S. and Inouye, M., Cell, 56: 709-717 (1989).

Inouye, S., Herzer, P.J. and Inouye, M., Proc. Natl. Acad. Sci, 87: 942-945 (1990).

Johnson M.S., McClure M.A., Feng D.F., Gray J., Doolittle R.F., Proc. Natl. Acad. Sci., USA, 83, 7648-7652 (1986).

Kaiser D., Ann. Rev. Genet, 20, 539-566 (1986).

Komano T., Franceschinti T., Inouye S., J. Mol. Biol., 196, 517-524 (1987).

Lampson B.C., Inouye M., Inouye S., Cell, 56, 701-707 (1989a).

Lampson B.C., Inouye S., Inouye M., "msDNA of Bacteria", Progress in Nucleic Acid Research and Molecular Biology, Vol. 40, pages 1 et seq.

Lampson B.C., Sun J., Hsu M.Y., Vallejo-Ramierez J., Inouye S., Inouye M., Science, 243, 1033-1038 (1989b).

Lampson B.C., Viswanathan M., Inouye M. and Inouye S., "Reverse Transcriptase from Escherichia coli Exists as a Complex with msDNA and is Able to Synthesize Double-stranded DNA", J. Bioll. Chem. 265: 8490- 8496 (1990)

Lim, D. and Maas, W.K., Cell, 56: 891-904 (1989).

Lim D., Maas W., Mol. Microbiol., 3 1141-1144 (1989).

Maniatis T., Fritsch E.F., Sambrook J., Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Laboratory, NY (1982).

Maruyama T., Gojobori T., Aota S., Ikemura T., Nuc. Acid Res. , 14, r151-r189 (1986).

Maxam A.M., Gilbert W., Meth. Enzymol., 65, 499 (1980).

Nargang F.E., Bell, J.B., Stohl L.L., Lambowitz A.M., Cell, 38, 441-453 (1984).

Perron Y.C., Vieria J., Messing J., Gene, 33, 103 (1985).

Ratner L., et al., Nature, 313, 277-283.

Rice N.R., Stephens R.M., Burny A., Gilden, R.V., Virology, 142, 357-377 (1985).

Romeo J.M., Emson B., Zusman D.R., Proc. Natl. Acad. Sci. USA, 83, 6332-6336 (1986).

Roth M.J., Tanese, N, Goff S.P., J. Biol. Chem., 260, 9326 (1985).

Ruskin B., Green M., Science, 229, 4274 (1987).

Saigo K., Kugimiya W., Matsuo Y., Inouye S., Yoshioka K., Yuki S., Nature, 312, 659-661 (1984).

Sanger F., Nicklen S., Coulson A.R., Proc. Natl. Acad. Sci., USA, 74, 5463-5467.

Seiki M., Hattori S., Hirayama Y., Yoshida M., Proc. Natl. Acad. Sci. USA, 80, 3618-3622 (1983).

Shinnick T.M., Lerner R.A., Sutcliffe J.G., Nature, 293, 543-548 (1981).

Stavenhagen, J.B., Robins D.M., Cell, 55, 247-254 (1988).

Southern E., J. Mol. Biol., 98, 503 (1975).

Tanese N., Roth M., Goff S.P., Proc. Natl. Acad. Sci. USA, 82, 4944 (1985).

Tanese N., Sodroski J., Haseltine W.A., Goff S.P., J. Virol., 59, 743 (1986).

Temin H.M., Cell, 21, 599-600 (1980).

Toh H., Hayashida H., Miyata T., Nature, 305, 827 (1983).

Varmus H.E., Nature, 314, 584-585 (1985).

Vieira J., Messing J., Gene, 19, 259-268 (1982).

Visawanathan M., Inouye M., Inouye S., J. Biol. Chem., 264, 13665-13671 (1989).

Voytas D.F., Ausbel F.M., Nature, 336, 242-244 (1988).

Weiss N., Teich H., Varmus H., Coffin J., RNA Tumor Viruses, Vol. 2, Cold Spring Harbor Laboratory (1985).

Yee T., Furuichi T., Inouye S., Inouye M., Cell, 38, 203-209 (1984).

Yuki S., Ishimaru S., Inouye S., Saigo K., Nucl. Acid Res., 14, 3017-3020 (1986).

ABSTRACT

The present invention relates to a prokaryotic reverse transcriptase enzyme. The enzyme is capable of synthesizing a hybrid DNA-RNA molecule called msDNA with the genes which synthesize the DNA and RNA portions of the molecule.

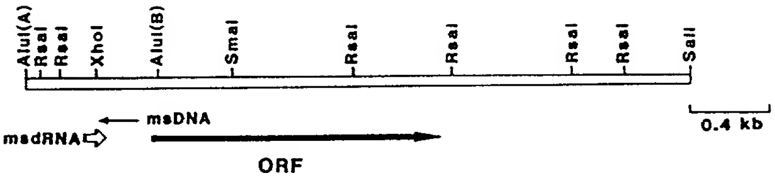

AluI(A)
RsaI
RsaI
XhoI
AluI(B)
SmaI
RsaI
RsaI
RsaI
RsaI
SalI
msDNA
msdRNA
ORF
0.4 kb

FIGURE 1

```
TCA TCC GCG CCG AGA CGC GCT CCT ACG TGC CGC GCG ACG CGG AGA GCC GCC TGG AGA CGC   60
TGT ACC GGG TTT CCC TGG ATG GTC AGG TGG TGG CGG TGG ACT GGG GCC CCC GCA CCC GCT  120
CGC CGC GTC ACC AGC GCC TCT GGT TCG ACT CGG ATG CGG AAG CCC CCG GAG CGT ACT TCG  180
CGC GCC TCG ACA AGT TCG CCC CTC ACG GCT ACA TCC ACC CGG CCT CGC CAT TGC TGT AAA  240
CCC TTC AAC CAC GGC TCG GCC CCC ACG CGC GGC CGG CAG GAC AGG TCC GAC GAA CAG ACC  300
                       a2                ⇨RNA
ACG ACG TGC GCT TCA CGC CCG AGC AGC CGA CAG AGG TCC GGA GTG CAT CAG CCT GAG CGC  360
TGC TGC ACG CGA AGT GCG GGC TCG TCG GCT GTC TCC AGG CCT CAC GTA GTC GGA CTC GCG

CTG GAG CGG CGG AGC GGC GTT GCG CCG CTC CGG TTG GAA TGC AGG ACA CTC TCC GCA AGG  420
CAG CTC GCC GCC TCG CCG CAA CGC GGC GAG GCC AAC CTT ACG TCC TGT GAG AGG CGT TCC

TAG CCT GTT GTT GGC TCT GTC CCT CCT AGG CAC TAG GGC CAG GGT GGG TAG CGG AGC CAA  480
ATC GGA CAA CAA CCG AGA CAG GGA GGA TCC GTG ATC CCG GTC CCA CCC ATC GCC TCG GTT

CGA CGC GAG CGC CGT TTA CGG ACG CCG CCC GTA GTG CCT AGG ACG CGA GAG CCG CTG AGG  540
GCT GCG CTC GCG GCA AAT GCC TGC GGC GGG CAT CAC GGA TCC TGC GCT CTC GGC GAC TCC
                                                   a1
CTA CCG TGC CCC AGG TAA GAT CGT GCT CCT TTC CCG CCC TCC CTC GAC TGC TCG CCC CAT  600
GAT GGC ACG GGG TCC ATT CTA GCA CGA GGA AAG GGC GGG AGG GAG CTG ACG AGC GGG GTA
                       DNA⇦
GTC CCG TCT TCC ATC GCC GCC CCC GCC CAA GGT GCA GAC ATG ACC GCC AGG CTG GAC CCG  660
                                                     M   T   A   R   L   D   P
                 AluI
TTC GTC CCC GCA GCT TCG CCG CAG GCC GTG CCC ACG CCC GAG CTC ACC GCT CCC TCG TCA  720
 F   V   P   A   A   S   P   Q   A   V   P   T   P   E   L   T   A   P   S   S
GAC GCG GCC GCG AAG CGT GAA GCC CGC CGG CTC GCC CAC GAA GCG TTG CTC GTC CGC GCG  780
 D   A   A   A   K   R   E   A   R   R   L   A   H   E   A   L   L   V   R   A
AAG GCC ATC GAC GAA GCG GGC GGC GCC GAC GAC TGG GTG CAG GCC CAG CTC GTC TCC AAG  840
 K   A   I   D   E   A   G   G   A   D   D   W   V   Q   A   Q   L   V   S   K
         *50
GGC CTC GCG GTG GAG GAC CTG GAC TTC TCC AGC GCC TCC GAG AAG GAC AAG AAG GCC TGG  900
 G   L   A   V   E   D   L   D   F   S   S   A   S   E   K   D   K   K   A   W
AAG GAG AAG AAG AAG GCC GAG GCC ACC GAG CGC CGC GCG CTG AAG CGT CAG GCG CAC GAG  960
 K   E   K   K   K   A   E   A   T   E   R   R   A   L   K   R   Q   A   H   E
                                                 *100
GCG TGG AAG GCC ACG CAC GTG GGC CAC CTG GGC GCG GGC GTG CAC TGG GCG GAG GAC CGC 1020
 A   W   K   A   T   H   V   G   H   L   G   A   G   V   H   W   A   E   D   R
                                                         SmaI
CTG GCC GAC GCG TTC GAC GTG CCC CAC CGC GAG GAG CGC GCC CGG GCC AAC GGC CTG ACG 1080
 L   A   D   A   F   D   V   P   H   R   E   E   R   A   R   A   N   G   L   T
GAG CTG GAC TCG GCG GAG GCG CTG GCC AAG GCC CTC GGC CTG AGC GTC TCC AAG CTG CGC 1140
 E   L   D   S   A   E   A   L   A   K   A   L   G   L   S   V   S   K   L   R
         *150
TGG TTC GCG TTC CAC CGC GAG GTG GAC ACG GCC ACG CAC TAC GTG AGC TGG ACG ATT CCG 1200
 W   F   A   F   H   R   E   V   D   T   A   T   H   Y   V   S   W   T   I   P
AAG CGG GAC GGC AGC AAG CGC ACG ATT ACG TCC CCC AAG CCT GAG CTC AAG GCA GCC CAG 1260
 K   R   D   G   S   K   R   T   I   T   S   P   K   P   E   L   K   A   A   Q
                                                 *200
CGC TGG GTC CTG TCC AAC GTC GTC GAG CGC CTG CCG GTC CAC GGC GCC GCG CAC GGC TTC 1320
 R   W   V   L   S   N   V   V   E   R   L   P   V   H   G   A   A   H   G   F
GTC GCC GGA CGC TCC ATC CTC ACC AAC GCG CTG GCC CAC CAG GGC GCC GAC GTG GTC GTG 1380
 V   A   G   R   S   I   L   T   N   A   L   A   H   Q   G   A   D   V   V   V
AAG GTG GAC CTC AAG GAC TTC TTC CCC TCC GTC ACC TGG CGC CGG GTC AAG GGC CTG TTG 1440
 K   V   D   L   K   D   F   F   P   S   V   T   W   R   R   V   K   G   L   L
         *250
CGC AAG GGC GGC CTG CGG GAG GGC ACG TCC ACG CTG CTG TCG CTC CTC TCC ACG GAA GCG 1500
 R   K   G   G   L   R   E   G   T   S   T   L   L   S   L   L   S   T   E   A
CCC CGC GAG GCG GTC CAG TTC CGC GGC AAG CTG CTG CAC GTG GCC AAG GGC CCC CGC GCG 1560
 P   R   E   A   V   Q   F   R   G   K   L   L   H   V   A   K   G   P   R   A
                                                 *300
CTG CCC CAG GGC GCG CCC ACG TCG CCG GGC ATC ACC AAC GCG CTG TGC CTG AAG CTG GAC 1620
 L   P   Q   G   A   P   T   S   P   G   I   T   N   A   L   C   L   K   L   D
AAC CGC CTG TCC GCG CTC GCG AAG CGG CTG GGC TTC ACC TAC ACC CGC TAC GCG GAC GAC 1680
 K   R   L   S   A   L   A   K   R   L   G   F   T   Y   T   R   Y   A   D   D
CTC ACC TTC TCG TGG ACG AAG GCG AAG CAG CCC AAG CCG CGC CGG ACC CAG CGT CCC CCG 1740
 L   T   F   S   W   T   K   A   K   Q   P   K   P   R   R   T   Q   R   P   P
         *350
GTC GCG GTG CTC CTG TCT CGC GTG CAG GAA GTG GTG GAG GCG GAG GGC TTC CGC GTG CAC 1800
 V   A   V   L   L   S   R   V   Q   E   V   V   E   A   E   G   F   R   V   H
CCG GAC AAG ACG CGC GTG GCG CGC AAG GGC ACG CGG CAG CGC GTG ACG GGC CTG GTC GTG 1860
 P   D   K   T   R   V   A   R   K   G   T   R   Q   R   V   T   G   L   V   V
                                                 *400
AAT GCG GCG GGC AAG GAC GCG CCG GCG GCC CGA GTC CCG CGC GAC GTC GTG CGC CAG CTC 1920
 N   A   A   G   K   D   A   P   A   A   R   V   P   R   D   V   V   R   Q   L
CGC GCC GCC ATC CAC AAC CGG AAG AAG GGC AAG CCG GGC CGC GAG GGC GAG TCG CTG GAG 1980
 R   A   A   I   H   N   R   K   K   G   K   P   G   R   E   G   E   S   L   E
CAG CTC AAG GGC ATG GCC GCC TTC ATC AAC ATG ACG GAC CCG GCC AAG GGC CGC GCC TTC 2040
 Q   L   K   G   H   A   A   F   I   N   N   T   D   P   A   K   G   R   A   F
         *450
CTG GCT CAG CTC ACG GAG CTG GAG TCC ACG GCG AGC GCG GCT CCG CAG GCG GAG TCA CCC 2100
 L   A   Q   L   T   E   L   E   S   T   A   S   A   A   P   Q   A   E
TCA CCC CGC GTC CGT CGC CGA GGT GCC GCG CCC CAG CAA CCC CGC ATT CAC GAA GTC CCT 2160
CAG CCG CCG CGG GTA C
```

FIGURE 2

```
HIV     RT   VKLKPGMDGPKVKQ  WPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKWR
HTLV1   RT   RPWARTPPKAPRNQ PVPFKPERLQALQHLVRKALEAGHIEPYTG  PGNNPVFPVKKA NGTWR
Ec-67   RT   NVLYRIGSDNQYTQFTIPKKGKGVRTISAPTDRL KDIQRRICDLLSDCRDEIFAIRKI SNNYS
Mx-162  RT   AFHREVDTATHYVSWTIPKRDGSKRTITSPKPEL KAAQR   WVLS    NVV ERLP VHGAA
                        o   oo●o     ooo  o   o o  oo     oo         o
                                             x  xx               x

HIV     RT   KLVDFRELNKRTQDFWEVQLGIPHPAGLKKK KSVTVLDVGDAYFSVPLDEDFRKYT       A
HTLV1   RT   FIHDLRATNSLTIDLSSSSPGPPDLSSLPTTLAHLQTIDLRDAFFQIPLPKQFQPYF       A
Ec-67   RT   FGFE RGKSIILNAYKHRGKQIILNIDLKDFFESFNFGRVRG YFLS  NQDF           L
Mx-162  RT   HGFV AGRSILTNALAHQGADVVVKVDLKDFFPSVTWRRVKGLLRKGGLREGTSTLLSLLSTEAP
              oo   o oo  oo  o o       o●oooo o    oo o
                       xx  x             x    xxx   x        x x

HIV     RT   FTIP SINNETPGIRYQYNVLPQGWKGSPAIFQS   SMTKILEPFKKQNPDIVIYQYMDDLYVG
HTLV1   RT   FTVP QQCNYGPGTRYAWKVLPQGFKNSPTLFEM   QLAHILQPIRQAFPQCTILQYMDDILLA
Ec-67   RT   LN PVVATTLAKAACYN GTLPQGSPCSPIISNLICNIMDMRLAKLAKKY GCTYSRYADDITI
Mx-162  RT   REAVQFRGKLLHVAKGP RALPQGAPTSPGITNALCLKLDKRLSALAKRL GFTYTRYADDLTF
                      o   o      ●●●● o ●● o o  o   o o●  ooo   o oo o●o●● o
                                           x       x x      x     x       x

HIV     RT   S DLEIGQHRTKIEELRQHLLRWGLTTP DKKHQKEP PFLWMGYELHPDKWTVQPIVLPE  KD
HTLV1   RT   S  PSHEDLLLLSEATMASLISHGLPVS ENKTQQTPGTIKFLGQIISPNHLTYDAVPTVPI RS
Ec-67   RT   STNKNTFPLEMATVQPEGVVLGKVLVKEIENSGFEINDSKTRLTYKTSRQEVT GLTVNRIVNID
Mx-162  RT   SWTKAKQPKPRRTQRPPVAVLLSRVQEVVEAEGFRVHPDKTRVARKGTRQRVT GLVVNAAGKDA
             ●  o   o    o  o   oo        o  oo     ooo   o oo o●  oo oo
                                 x        x                          xx

HIV     RT   SWTVNDIQKLVGKLNWASQIYP
HTLV1   RT   RWALPELQALLGEIQWVSKGTP
Ec-67   RT   RCYYKKTRALAHALYRTGE YK
Mx-162  RT   PAARVPRDVVRQLRAAIHN RK
                                  o
               x
```

FIGURE 3

A

```
Mx-162     18 PTPELTAPSSDAAAKREARRLAHEALLVRAKAIDEAGGADDWVQAQLVSKGLAVEDLD-FSSASEKDKKA-WKEKK   91

Mo-MLV   1070 PDPDMTRVTNSPSLQAHLQALYLVQHEVW-RPL-AAAYQEQ-LDRPVVPHPYRVGDTVWVRRHQTKNLEPRWKGPY 1142
              o o  o              o      o       o           o     o o         o     oo


Mx 162     92 KAEATERRALKRQAHEAW-KATHVGHLGAGVHWAEDRL       128

Mo-MLV   1143 TVLLTTPTALKVDGIAAWIHAAHVKAADPGGG-PSSRL      1179
                  o   ooo     oo  o oo     o      oo
```

B

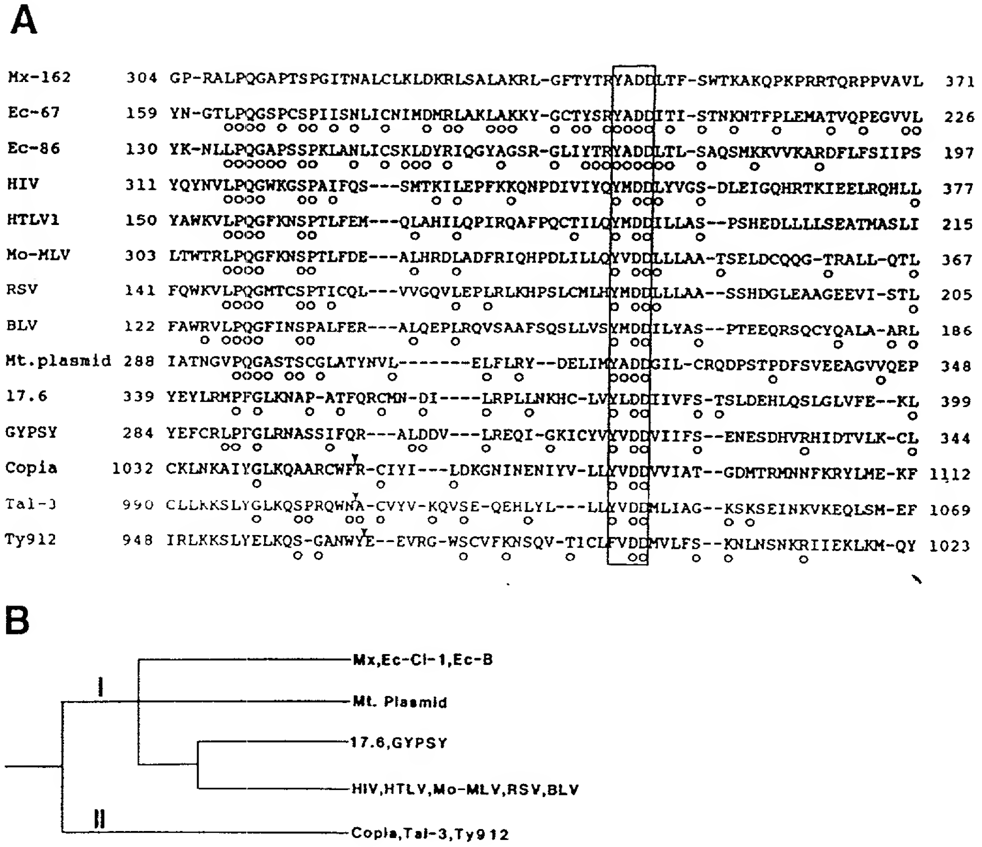
```
Mx-162    411 GKDAPAARVPRDVVRQLRAAIHNRKKGKPGREGESLEQLKGMAAFIHMTD-PAKGRAF-LAQLTELESTASAAPQAE 485

HIV       396 GKEGHSARQCR-APR--RQGC--WKCGKPGHIMTNCPD-R-QAGFLGLGPWGKKPRNFPVAQVPQ-GLTPTAPP    461
              oo    oo  o   o  o      o oooo            o o        o o o  oo      o  o o
```

Figure 4. Sequence Similarity of the msDNA-Mx162 Reverse Transcriptase with Other Retroelements

(A) Sequence similarity of the region from residues 18 to 128 of the msDNA Mx162 RT (see Figure 2) with a carboxy-terminal region of integration protein of Moloney murine leukemia virus (M-MuLV) (residues 1070 to 1179, Shinnick et al . 1981)

(B) Comparison of the sequence from residues 411 to 485 of the msDNA-Mx162 RT (see Figure 2) with the sequence from residues 396 to 461 of the *gag* protein of human immunodeficiency virus (HIV, Ratner et al , 1985)

FIGURE 4



```
Mx-162       304  GP-RALPQGAPTSPGITNALCLKLDKRLSALAKRL-GFTYTRYADDLTF-SWTKAKQPKPRRTQRPPVAVL   371
Ec-67        159  YN-GTLPQGSPCSPIISNLICNIMDMRLAKLAKKY-GCTYSRYADDITI-STNKNTFPLEMATVQPEGVVL   226
Ec-86        130  YK-NLLPQGAPSSPKLANLICSKLDYRIQGYAGSR-GLIYTRYADDLTL-SAQSMKKVVKARDFLFSIIPS   197
HIV          311  YQYNVLPQGWKGSPAIFQS---SMTKILEPFKKQNPDIVIYQYMDDLYVGS-DLEIGQHRTKIEELRQHLL   377
HTLV1        150  YAWKVLPQGFKNSPTLFEM---QLAHILQPIRQAFPQCTILQYMDDILLAS--PSHEDLLLLSEATMASLI   215
Mo-MLV       303  LTWTRLPQGFKNSPTLFDE---ALHRDLADFRIQHPDLILLQYVDDLLLAA-TSELDCQQG-TRALL-QTL   367
RSV          141  FQWKVLPQGMTCSPTICQL---VVGQVLEPLRLKHPSLCMLHYMDDLLLAA--SSHDGLEAAGEEVI-STL   205
BLV          122  FAWRVLPQGFINSPALFER---ALQEPLRQVSAAFSQSLLVSYMDDILYAS--PTEEQRSQCYQALA-ARL   186
Mt.plasmid   288  IATNGVPQGASTSCGLATYNVL-------ELFLRY--DELIMYADDGIL-CRQDPSTPDFSVEEAGVVQEP   348
17.6         339  YEYLRMPFGLKNAP-ATFQRCMN-DI----LRPLLNKHC-LVYLDDIIVFS-TSLDEHLQSLGLVFE--KL   399
GYPSY        284  YEFCRLPFGLRNASSIFQR---ALDDV---LREQI-GKICYVYVDDVIIFS--ENESDHVRHIDTVLK-CL   344
Copia       1032  CKLNKAIYGLKQAARCWFR-CIYI---LDKGNINENIYV-LLYVDDVVIAT--GDMTRMNNFKRYLME-KF  1112
Tal-3        990  CLLKKSLYGLKQSPRQWNA-CVYV-KQVSE-QEHLYL---LLYVDDMLIAG--KSKSEINKVKEQLSM-EF  1069
Ty912        948  IRLKKSLYELKQS-GANWYE--EVRG-WSCVFKNSQV-TICLFVDDMVLFS--KNLNSNKRIIEKLKM-QY  1023
```

FIGURE 5

FIGURE 6

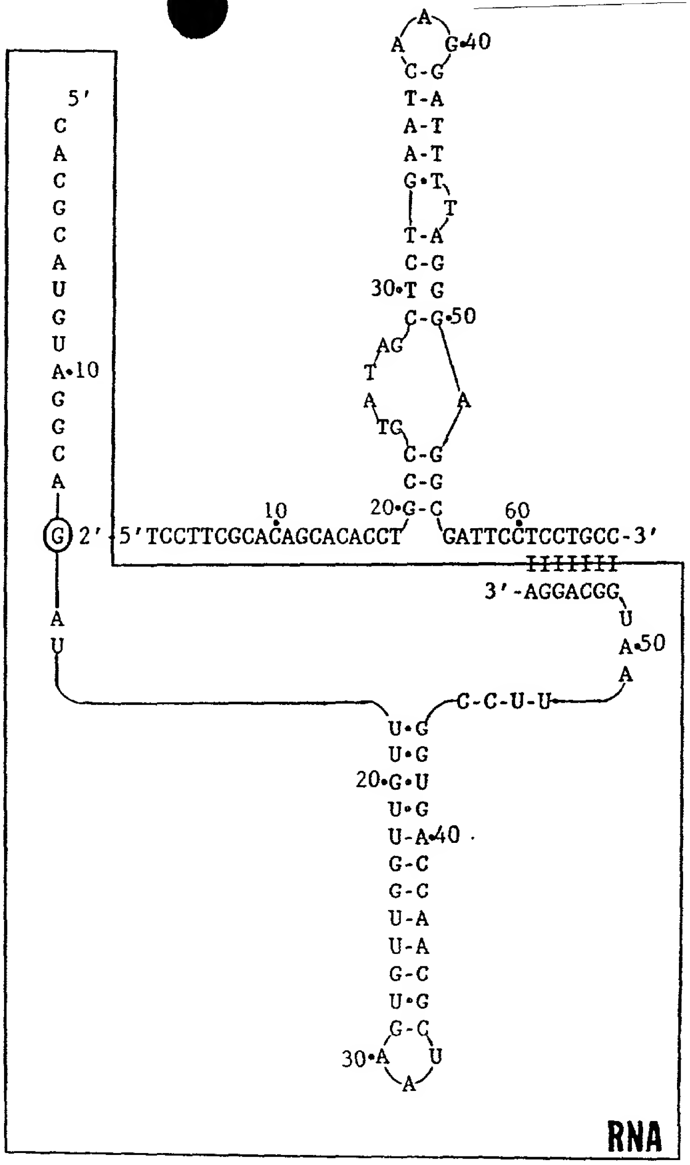


FIGURE 7

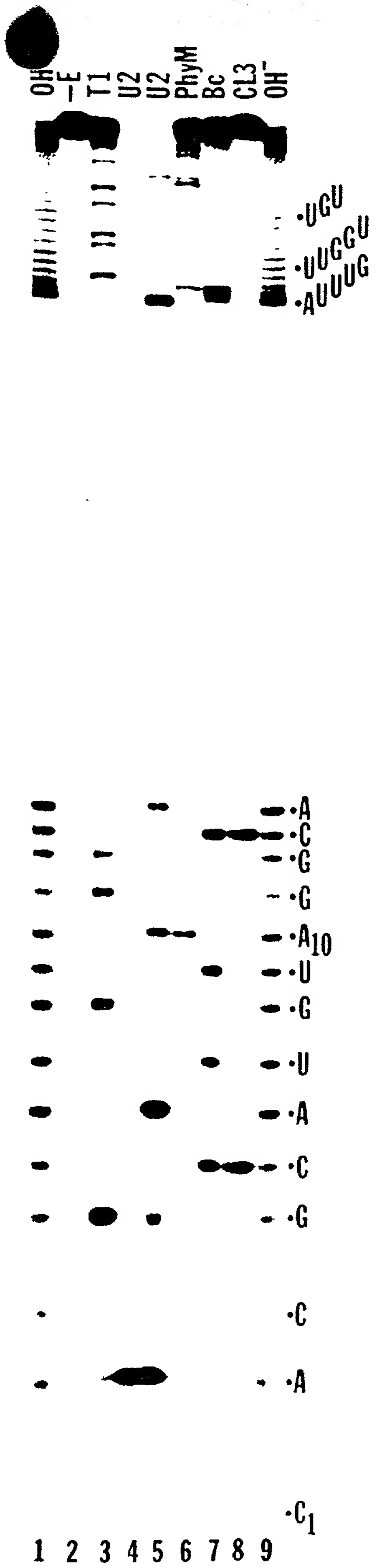
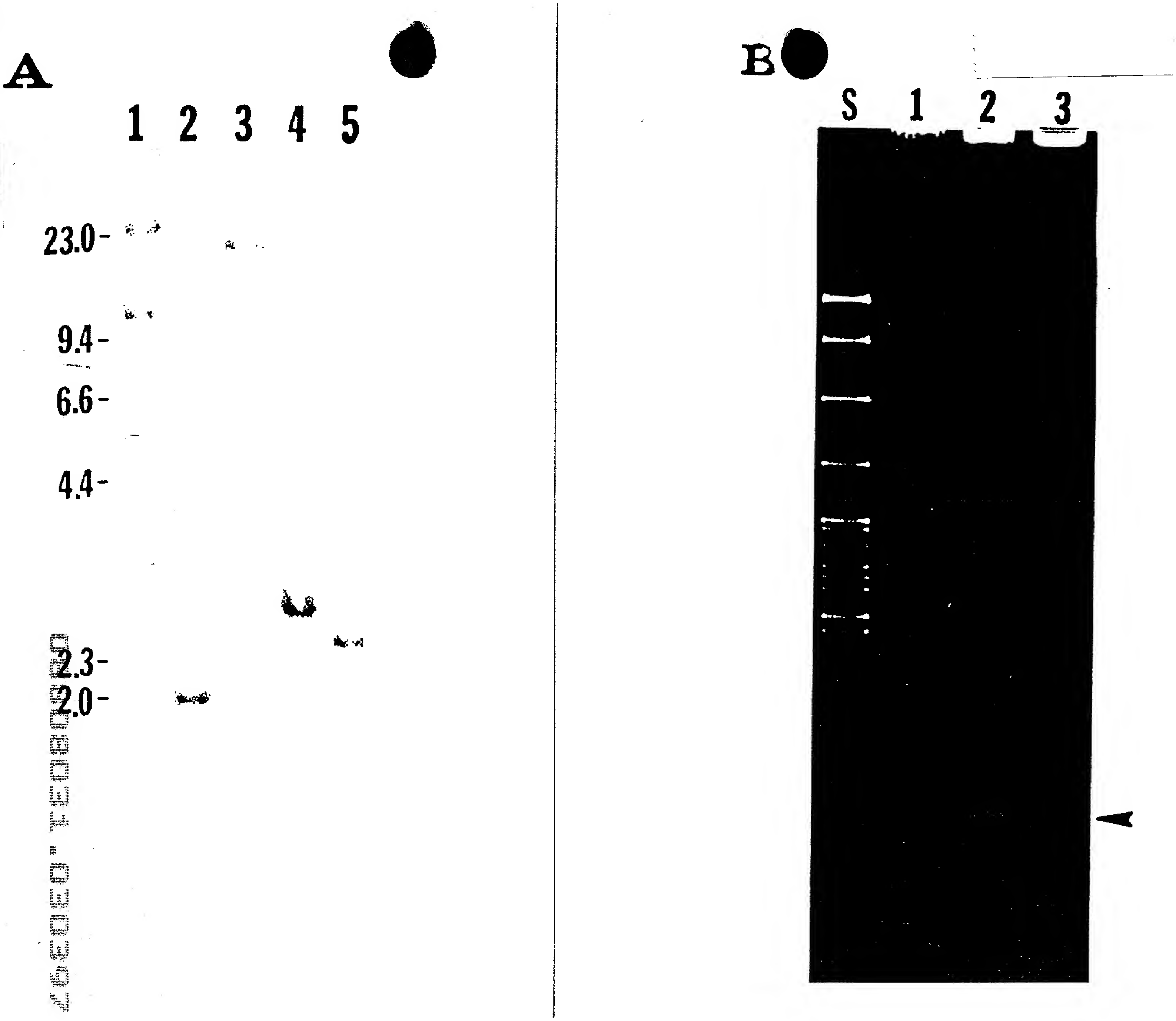

OH⁻
–E
T1
U2
U2
PhyM
Bc
CL3
OH⁻
UGU
UUGGU
AUUUG
A
C
G
G
A10
U
G
U
A
C
G
C
A
C1
1 2 3 4 5 6 7 8 9

FIGURE 8



FIGU

FIGURE 9

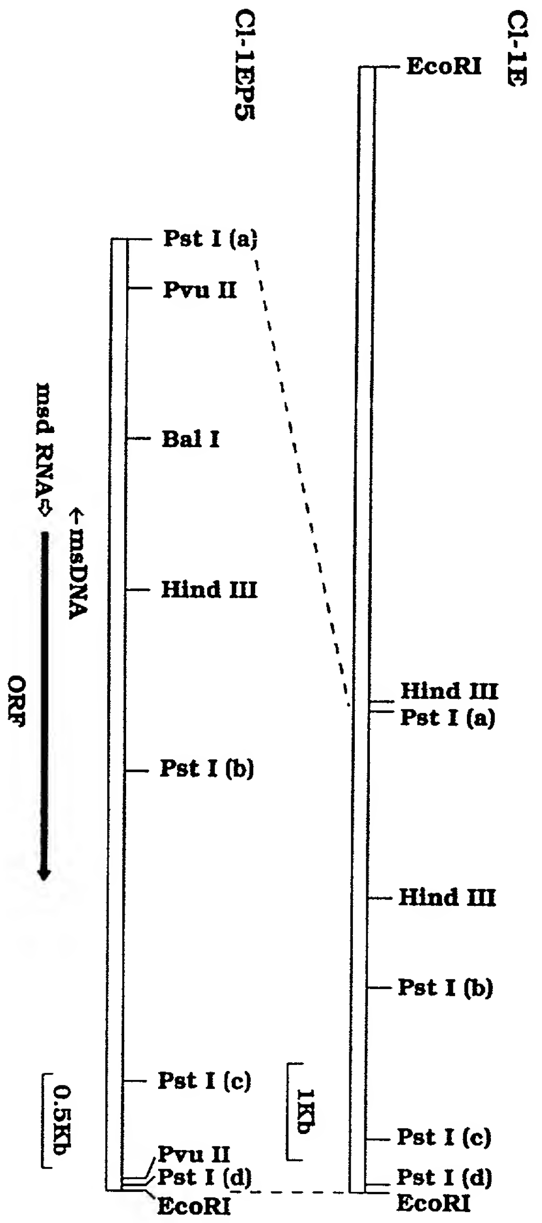

Cl-1EP5
Cl-1E
EcoRI
Pst I (a)
Pvu II
Bal I
Hind III
Pst I (b)
Pst I (c)
Pvu II
Pst I (d)
EcoRI
←msDNA
msd RNA⇨
ORF
0.5Kb
1Kb
Hind III
Pst I (a)
Hind III
Pst I (b)
Pst I (c)
Pst I (d)
EcoRI

FIGURE 10

```
                                             -35                      -10
TCG CCA T[illegible] CGA TTT TCA CTT CCT TGA CAG TGC ATG ACT ATG CTG CAT GAA A[illegible]
GCA TGA TCG ATT GAG GAT CGT CTT TGC TCA GAT CCG CCA GAA CTG GCC GGC TTT TGC TCA  120
TCT CAT GCA TGT GCA TGA AAA CCA CTG CAT AAA CCG GGC AGG CCT GCC GGG GAT ACC ACC  180
GCG CGC TAT CAC CGA AAA TAG CCA AAA TAC TTC TGG AAA ACA GAA AGT TCA AGT GAT ATG  240
        RNA      a2
TTC ATA AAC ACG CAT GTA GGC AGA TTT GTT GGT TGT GAA TCG CAA CCA GTG GCC TTA ATG  300
AAG TAT TTG TGC GTA CAT CCG TCT AAA CAA CCA ACA CTT AGC GTT GGT CAC CGG AAT TAC

GCA GGA GGA ATC GCC TGC GTA AAA TGC TTG ATT CAG AGC TAT ACG GCA GGT GTG CTG TGC  360
CGT CCT CCT TAG CGG AGG GAT TTT AGG AAC TAA GTC TCG ATA TGC CGT CCA CAC GAC ACG

GAA GGA GTG CCT GCA TGC GTT TCT CCT TCG CCT TTT TTC CTC TGG GAT GAA GAA GAA ATG  420
CTT CCT CAC GGA CGT ACG CAA AGA GGA ACC GGA AAA AAG GAG ACC CTA CTT CTT CTT TAC
     DNA          a1                                                        M
ACA AAA ACA TCT AAA CTT GAC GCA CTT AGG GCT GCT ACT TCA CGT GAA GAC TTG GCT AAA  480
 T   K   T   S   K   L   D   A   L   R   A   A   T   S   R   E   D   L   A   K
ATT TTA GAT ATT AAG TTG GTA TTT TTA ACT AAC GTT CTA TAT AGA ATC GGC TCG GAT AAT  540
 I   L   D   I   K   L   V   F   L   T   N   V   L   Y   R   I   G   S   D   N
CAA TAC ACT CAA TTT ACA ATA CCG AAG AAA GGA AAA GGG GTA AGG ACT ATT TCT GCA CCT  600
 Q   Y   T   Q   F   T   I   P   K   K   G   K   G   V   R   T   I   S   A   P
                                *50
ACA GAC CGG TTG AAG GAC ATC CAA CGA AGA ATA TGT GAC TTA CTT TCT GAT TGT AGA GAT  660
 T   D   R   L   K   D   I   Q   R   R   I   C   D   L   L   S   D   C   R   D
GAG ATC TTT GCT ATA AGG AAA ATT AGT AAC AAC TAT TCC TTT GGT TTT GAG AGG GGA AAA  720
 E   I   F   A   I   R   K   I   S   N   N   Y   S   F   G   F   E   R   G   K
                                                                       *100
TCA ATA ATC CTA AAT GCT TAT AAG CAT AGA GGC AAA CAA ATA ATA TTA AAT ATA GAT CTT  780
 S   I   I   L   N   A   Y   K   H   R   G   K   Q   I   I   L   N   I   D   L
                     H
AAG GAT TTT TTT GAA AGC TTT AAT TTT GGA CGA GTT AGA GGA TAT TTT CTT TCC AAT CAG  840
 K   D   F   F   E   S   F   N-  F   G   R   V   R   G   Y   F   L   S   N   Q
GAT TTT TTA TTA AAT CCT GTG GTG GCA ACG ACA CTT GCA AAA GCT GCA TGC TAT AAT GGA  900
 D   F   L   L   N   P   V   V   A   T   T   L   A   K   A   A   C   Y   N   G
                                *150
ACC CTC CCC CAA GGA AGT CCA TGT TCT CCT ATT ATC TCA AAT CTA ATT TGC AAT ATT ATG  960
 T   L   P   Q   G   S   P   C   S   P   I   I   S   N   L   I   C   N   I   M
GAT ATG AGA TTA GCT AAG CTG GCT AAA AAA TAT GGA TGT ACT TAT AGC AGA TAT GCT GAT 1020
 D   M   R   L   A   K   L   A   K   K   Y   G   C   T   Y   S   R   Y   A   D
                                                                       *200
GAT ATA ACA ATT TCT ACA AAT AAA AAT ACA TTT CCG TTA GAA ATG GCT ACT GTG CAA CCT 1080
 D   I   T   I   S   T   N   K   N   T   F   P   L   E   M   A   T   V   Q   P
GAA GGG GTT GTT TTG GGA AAA GTT TTG GTA AAA GAA ATA GAA AAC TCT GGA TTC GAA ATA 1140
 E   G   V   V   L   G   K   V   L   V   K   E   I   E   N   S   G   F   E   I
AAT GAT TCA AAG ACT AGG CTT ACG TAT AAG ACA TCA AGG CAA GAA GTA ACG GGA CTT ACA 1200
 N   D   S   K   T   R   L   T   Y   K   T   S   R   Q   E   V   T   G   L   T
                                *250
GTT AAC AGA ATC GTT AAT ATT GAT AGA TGT TAT TAT AAA AAA ACT CGG GCG TTG GCA CAT 1260
 V   N   R   I   V   N   I   D   R   C   Y   Y   K   K   T   R   A   L   A   H
GCT TTG TAT CGT ACA GGT GAA TAT AAA GTG CCA GAT GAA AAT GGT GTT TTA GTT TCA GGA 1320
 A   L   Y   R   T   G   E   Y   K   V   P   D   E   N   G   V   L   V   S   G
                                                                     *300
GGT CTG GAT AAA CTT GAG GGG ATG TTT GGT TTT ATT GAT CAA GTT GAT AAG TTT AAC AAT 1380
 G   L   D   K   L   E   G   M   F   G   F   I   D   Q   V   D   K   F   N   N
ATA AAG AAA AAA CTG AAC AAG CAA CCT GAT AGA TAT GTA TTG ACT AAT GCG ACT TTG CAT 1440
 I   K   K   K   L   N   K   Q   P   D   R   Y   V   L   T   N   A   T   L   H
GGT TTT AAA TTA AAG TTG AAT GCG CGA GAA AAA GCA TAT AGT AAA TTT ATT TAC TAT AAA 1500
 G   F   K   L   K   L   N   A   R   E   K   A   Y   S   K   F   I   Y   Y   K
                                *350
TTT TTT CAT GGC AAC ACC TGT CCT ACG ATA ATT ACA GAA GGG AAG ACT GAT CGG ATA TAT 1560
 F   F   H   G   N   T   C   P   T   I   I   T   E   G   K   T   D   R   I   Y
TTG AAG GCT GCT TTG CAT TCT TTG GAG ACA TCA TAT CCT GAG TTG TTT AGA GAA AAA ACA 1620
 L   K   A   A   L   H   S   L   E   T   S   Y   P   E   L   F   R   E   K   T
                                                                     *400
GAT AGT AAA AAG AAA GAA ATA AAT CTT AAT ATA TTT AAA TCT AAT GAA AAG ACC AAA TAT 1680
 D   S   K   K   K   E   I   N   L   N   I   F   K   S   N   E   K   T   K   Y
                                  P
TTT TTA GAT CTT TCT GGG GGA ACT GCA GAT CTG AAA AAA TTT GTA GAG CGT TAT AAA AAT 1740
 F   L   D   L   S   G   G   T   A   D   L   K   K   F   V   E   R   Y   K   N
AAT TAT GCT TCT TAT TAT GGT TCT GTT CCA AAA CAG CCA GTG ATT ATG GTT CTT GAT AAT 1800
 N   Y   A   S   Y   Y   G   S   V   P   K   Q   P   V   I   M   V   L   D   N
                                *450
GAT ACA GGT CCA AGC GAT TTA CTT AAT TTT CTG CGC AAT AAA GTT AAA AGC TGC CCA GAC 1860
 D   T   G   P   S   D   L   L   N   F   L   R   N   K   V   K   S   C   P   D
GAT GTA ACT GAA ATG AGA AAG ATG AAA TAT ATT CAT GTT TTC TAT AAT TTA TAT ATA GTT 1920
 D   V   T   E   M   R   K   M   K   Y   I   H   V   F   Y   N   L   Y   I   V
                                                                     *500
CTC ACA CCA TTG AGT CCT TCC GGC GAA CAA ACT TCA ATG GAG GAT CTT TTC CCT AAA GAT 1980
 L   T   P   L   S   P   S   G   E   Q   T   S   M   E   D   L   F   P   K   D
ATT TTA GAT ATC AAG ATT GAT GGT AAG AAA TTC AAC AAA AAT AAT GAT GGA GAC TCA AAA 2040
 I   L   D   I   K   I   D   G   K   K   F   N   K   N   N   D   G   D   S   K
ACG GAA TAT GGG AAG CAT ATT TTT TCC ATG AGC GTT GTT AGA GAT AAA AAC CGC AAA ATA 2100
 T   E   Y   G   K   H   I   F   S   M   R   V   V   R   D   K   K   R   K   I
                                *550
GAT TTT AAG GCA TTT TGT TGT ATT TTT GAT GCT ATA AAA GAT ATA AAG GAA CAT TAT AAA 2160
 D   F   K   A   F   C   C   I   F   D   A   I   K   D   I   K   E   H   Y   K
TTA ATG TTA AAT AGC TAA TCA ACA CCC CTA ACG TTA TCA ACG CTA ACG CTG ATT TTT CCT 2220
 L   M   L   N   S   -
TAA AAT TTA TAT GGT TTG AAT TGT AAT ATA TTA TCT TCA AGC CAT TTA TTT AAT TCG TGC 2280
ATC CTT TTC TGT AAG GGT ATT AAT TCG TTC CTC ACA AAC ACT AAA CTC GCT TTT TCG ACA 2340
TCC CCA AAC CCC CCT AAC ATT ATT CGG CAT AAT CCC CAT CAT TTG CGG TGG CAC ACG ATG 2400
CGC TGC CAT CAT GTC ATC GCG CC
```

FIGURE 11

```
V    RT    VKLKPGMDGPKVKQ  WPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKWR 239
LV1  RT    RPWARTPPKAPRNQ PVPFKPERLQALQHLVRKALEAGHIEPYTG   PGNNPVFPVKKA NGTWR  75
DNA  RT    NVLYRIGSDNQYTQFTIPKKGKGVRTISAPTDRL KDIQRRICDLLSDCRDEIFAIRKI SNNYS  94
             +  o       ●   ● o               +                 ●++ ●  +

V    RT    KLVDFRELNKRTQDFWEVQLGIPHPAGLKKK KSVTVLDVGDAYFSVPLDEDFRKYTAFTIP SI 302
LV1  RT    FIHDLRATNSLTIDLSSSSPGPPDLSSLPTTLAHLQTIDLRDAFFQIPLPKQFQPYFAFTVP QQ 139
DNA  RT    FGFE RGKSIILNAYKHRGKQIILNIDLKDFFESFNFGRVRG YFLS  NQDF    LLN PVVA 150
           o    ●               +     ●+    +     +o  +●      +●        ●

V    RT    NNETPGIRYQYNVLPQGWKGSPAIFQS   SMTKILEPFKKQNPDIVIYQYMDDLYVGS DLEIG 363
LV1  RT    CNYGPGTRYAWKVLPQGFKNSPTLFEM   QLAHILQPIRQAFPQCTILQYMDDILLAS  PSHE 199
DNA  RT    TTLAKAACYN GTLPQGSPCSPIISNLICNIMDMRLAKLAKKY GCTYSRYADDITI STNKNTF 212
                   ●    ●●●●   ●● +       +   o    +    oo   ●  ●●o   ●

V    RT    QHRTKIEELRQHLLRWGLTTP DKKHQKEP PFLWMGYELHPDKWTVQPIVLPE  KDSWTVNDI 424
LV1  RT    DLLLLSEATMASLISHGLPVS ENKTQQTPGTIKFLGQIISPNHLTYDAVPTVPI RSRWALPEL 262
DNA  RT    PLEMATVQPEGVVLGKVLVKEIENSGFEINDSKTRLTYKTSRQEVT GLTVNRIVNIDRCYYKKT 276
            o           +   ●    oo           o +  o    ●    +     +o

                         ▼
V    RT    QKLVGKLNWASQIYPGIK  VRQLCKLLRGTKALTEVIPLTEEAELELAENREILKEPVHGVYYD 487
LV1  RT    QALLGEIQWVSKGTPTLRQPLHSLYCALQRHTDPRDQIYLNPSQVQSLVQLRQALSQNCRSRLVQ 327
DNA  RT    RALAHALYRTGE YKVPDE   NGV  LVSGGLDKLEGMFGFIDQVDKFNNIKKKLNKQ PDRYVL 335
            o●   +      +            +  +  o  +      oo          ●     o+o

V    RT    PSKDLIA EIQKQGQGQWTYQIYQE PFKNLKTGKYARMRGAHTNDVKQLTEAVQKITT       544
LV1  RT    TLPLLGAIMLTLTGTTTVVFQSKEQWPLVWLHAPLPHTSQCPWGQLLASAVLLLDKYTLQSY GL 391
DNA  RT    TNATLHGFKLKL NAREKAY SKFIY YKFFHGNTCPTIITEGKTDRIYLKAALHSLET SYPEL 396
           o   ●    o o       + oo     +  o     o      +   +  +o    + oo  o

V    RT      ESIVIWGKTPKFKLPIQKETWETWWTEYWQATWI       PE WEFV  NTPPL  VKLWYQ 595
LV1  RT    LCQTIHHNISTQTFNQFIQTSDHPSVPILLHHSHRFKNLGAQTGELWNTFLKTAAPLAPVKALMP 456
DNA  RT    FREKTDSKKKEINLNIFKSNEKTKYFLDLSGGTADLKKFVERYKNNYASYYGSV PKQPVIMVLD 460
             +      +    o o   +       o       o                 ●  o●

                ▼
V    RT    LE KEPIV   GAETFYVDGAANRETKLGKAGYVTNKGRQK VV PLTNTTNQ  KTELQAIYLA 652
LV1  RT    VFTLSP VIINTAPCLFSDGSTSRAAYILWDKQILSQRS FP LPPPHKSA Q RAELLGLLHGL 516
DNA  RT    NDTG PSDLLN FLRNKVKSCPDDVTEMRKMKYIHVFYNLYIVLTPLSPSGEQTSMEDLFPKDIL 523
             o  ●    o      +       +  + o+o        +o ●+  o  ●   o ●     o

V    RT    LQDS GLE  VNIVTDSQYAL  QIIQA    QPDKSESELVNQIIEQLIKKEKVYLAWVPAHKG 708
LV1  RT    SSAR SWR CLNIFLDSKYLYHYLRTLALGTFQGRSSQAPFQA LLPRLLSRKVVYLHHVRSHTN 578
DNA  RT    DIKIDGKKFNKNNDGDSKTEYGKHI      FSMR        VV RDKKRKIDFKAFCCIFDA 572
                +     ●   ●●o  o  +      o  o            o  +oo    +

V    RT    IGGNEQVDKLVSAG                                                    722
LV1  RT    LPDPISRLNALTDA                                                    592
DNA  RT    IKDIKEHYKLMLNS                                                    586
           + o     ++
```

FIGURE 12

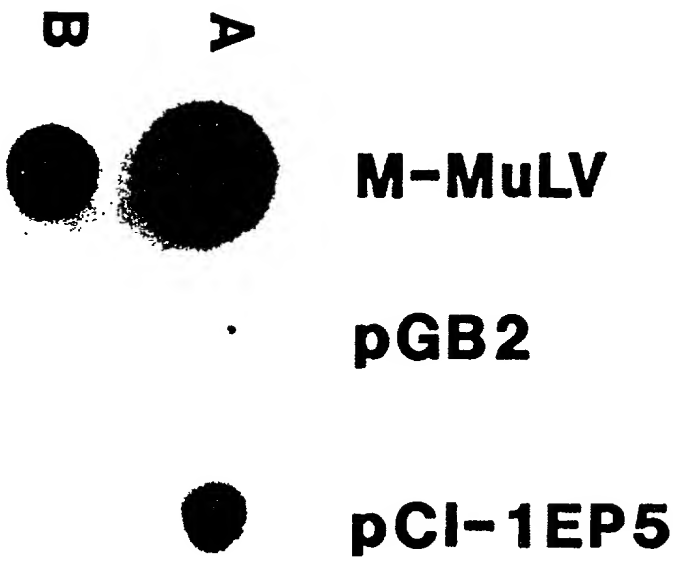


FIGURE 13

```
                                                         K
                              h hh h  K      hR  h       R
Sa163    165   RWFSFHREVD TGTHYQTWEI PKRDGG--KR TLTAPKRELK AVQRWVLANV VERLPVH--- -----GAAHG
Mx162    167   RWFAFHREVD TATHYVSWTI PKRDGS--KR TITSPKPELK AAQRWVLSNV VERLPVH--- -----GAAHG
Mx65     136   RHYSIHRPRE RVRHYVTFAV PKRSGG--VR LLHAPKRRLK ALQRRMLALL VSKLPVS--- -----PQAHG
Ec67      29   FLTNVLYRIG SDNQYTQFTI PKKGKG--VR TISAPTDRLK DIQRRICDLL SDCRDEIFAI RKISNNYSFG
Ec86      34   VETLRLLIYT ADFRYRIYTV EKKGPEKRMR TIYQPSRELK ALQGWVLRNI LDKLSSS--- -----PFSIG
Ec73      14   TKGFASEVMR SPEPPKKWDI AKKKGG--MR TIYHPSSKVK LIQYWLMNNV FSKLPMH--- -----NAAYA
Ec107     25   IQRLHALSNH AGRHYRRIIL SKRHGG--QR LVLAPDYLLK TVQRNILKNV LSQFPLS--- -----PFATA
                                 ● ●  ●●      ●  ●  ●   ●●   ●  ●   ●  ●
Consensus      ---------------Y--h-h -KR-GG  -R Th--P---LK -hQR-hL--h hp-LPhp         -hA-G
                                       K
                               1                            2

                                              A F
                                      h  hDh  G Y  h
Sa163    225   FVAGRSILTN ALAH--QGAD VVVKVDMKDF FPSVTWPRVK GLLRKGGLPE NLATLLALLS TEAPREVVRF
Mx162    227   FVAGRSILTN ALAH--QGAD VVVKVDLKDF FPSVTWRRVK GLLRKGGLRE GTSTLLSLLS TEAPREAVQF
Mx65     196   FVPGRSIKTG AAPH--VGRR VVLKLDLKDF FPSVTFARVR GLLKALGYGY PVAATLAVLM TESERQPVEL
Ec67      97   FERGKSIILN AYKH--RGKQ IILNIDLKDF FESFNFGRVR GYFLSNQDFL LNPVVATTLA ----------
Ec86      96   FEKHQSILNN ATPH--IGAN FILNIDLEDF FPSLTANKVF GVFHSLGYNR LISSVLT--- ----------
Ec73      74   FVKNRSIKSN ALLHAESKNK YYVKIDLKDF FPSIKFTDFE YAFTRYRDRI EFTTEYDLEL LQLIKT----
Ec107     85   YRPGCPIVSN AQPH--CQQP QILKLDIENF FDSISWLQVW RVFRQAQLPR NVVTMLT--- ----------
               ●     ●    ●  ●        ●● ●●●  ● ● ●●●●  ●   ●●             ●
Consensus      F--GRSIhpN A--H  -G-- hhhKhDhKDF FPShph-RVp Ghh------- ---Shh----
                   K                                   K                 T
                                                3

                             hPQG    pP hh   h                      h   Y DDhhh
Sa163    293   RGETLYVAKG PRALPQGAPT SPALTNALCL RLDKRLSALS ---KRLGFTY TRYADDLTFS WRRAKKSRQK
Mx162    285   PRELLHVAKG PRALPQGAPT SPGITNALCL KLDKRLSALA ---KRLGFTY TRYADDLTFS WTKAKQPKPR
Mx65     264   EGILFHVPVG PRVCVQGAPT SPALCNAVLL RLDRRLAGLA ---RRYGYTY TRYADDLTFS GDDVTA----
Ec67     155   -----KAACY NGTLPQGSPC SPIISNLICN IMDMRLAKLA ---KKYGCTY SRYADDITIS TNKNTFPLEM
Ec86     151   -----KICCY KNLLPQGAPS SPKLANLICS KLDYRIQGYA ---GSRGLIY TRYADDLTLS AQSMKK----
Ec73     140   -----ICFIS DSTLPIGFPT SPLIANFVAR ELDEKLTQKL NAIDKLNATY TRYADDIIVS TNMKGA----
Ec107    140   -----WICCY NDALPQGAPT SPAISNLVMR RFDERIGEWC ---QARGITY TRYCDDMTFS GHFNAR----
                                ● ●  ●● ● ●●●    ●● ●●              ● ●●● ●●● ●●  ●
Consensus           -h--- --hLPQGAPT SP-h-Nhh-- KLDpRL--h-    pp-GhTY TRYADDhThS -pp---
                                                R
                                         4                                    5

                                        Gh h  c K  h          hLG  h
Sa163    360   ELPLADAPVA LLLARVKGVL EAEGFTLHPD KTRVQRK--G SRQRVTGLVV       407
Mx162    362   --RTQRPPVA VLLSRVQEVV EAEGFRVHPD KTRVARK--G TRQRVTGLVV       407
Mx65     327   --------LE RVRALAARYV QEEGFEVNRE KTRVQRR--G GAQRVTGVTV       366
Ec67     217   --ATVQPEGV VLGKVLVKEI ENSGFEINDS KTRLTYK--T SRQEVTGLTV       262
Ec86     209   --------VV KARDFLFSII PSEGLVINSK KTCISGP--R SQRKVTGLVI       248
Ec73     201   -------SKL ILDCFKRTMK EIGPDFKINI KKFKICSASG GSIVVTGLKV       243
Ec107    198   ---------- QVKNKVCGLL AELGLSLNKR KGCLIAA--C KRQQVTGIVV       235
                           ●                    ●             ●●●● ●
Consensus              -- -h---h-phh p-pGhphppp KT-h--p  - ppQpVTGL-V
                                      6                        7
```

FIGURE 14

```
CT   GCCCCGCCTCCGAGGACGCGCTCGCGGCCCCGGCGGCGGGGCGGACGG   CG   60

GCGGCCCACGGAGACGCTTGACCCGGAGACGACGAATGACGATAACGGCAGGTGCTCTC  120

                  a2         ⇨RNA
CGGAGAGCCCAGCGCTCGCAGATGAGCCATGAGTACCGCGGTGTTTCGCCCGCGGGGTGT  180
CCCTCTCCGGTCCCGAGCGTCTACTCGGTACTCATGGCGCCACAAAGCGGCGCCCCCACA

TCTGTCCCCATCTCTTCGCCACGGTCCCAGCGTACGCAACGCAGGAGCCCCGGGTCCAA  240
AGACAGGGGTAGAGAAGCGGTCCCAGGGTCGCATGCGTTGCGTCCTCGGGGCCCAGGTT

                  a1                         AluI
CGCCTCGCAGGTCGTCCCCTGCCCTCTTCCGGAGCACCATGAGCTGGTTCGACACCACCC  300
GCGGAGCGTCCAGCAGGGGACCGGAGAAGGCCTCGTGGTACTCGACCAAGCTGTGGTGGG
     ⇦DNA                               M  S  W  F  D  T  T  L

TCTCCCGGCTCAAGGGGTTGTTCAGCCGTCCCGTGACACGAAGCACCACCGGGCTGGACC  360
 S  R  L  K  G  L  F  S  R  P  V  T  R  S  T  T  G  L  D  V

TGCCGCTGGATGCCCACGGACCTCCCCAGGACGTCGTGACGGAGACGGTCTCCACGTCGG  42C
 P  L  D  A  H  G  R  P  Q  D  V  V  T  E  T  V  S  T  S  G

GCCCCCTGAAGCCAGGGCACCTGCGACAGGTCCGCCGGGATGCGCGGCTGCTCCCCAAGG  480
 P  L  K  P  G  H  L  R  Q  V  R  R  D  A  R  L  L  P  K  G
   *50

GCGTCCGCCGCTACACCCCGGGCCGGAAGAAGTGGATGGAGGCCGCCGAGGCCCGGCGGC  540
 V  R  R  Y  T  P  G  R  K  K  W  M  E  A  A  E  A  R  R  L

TGTTCTCCGCCACGCTGCGCACGCGGAACCGGAACCTGAGGGACTTGCTGCCCGACGAGG  600
 F  S  A  T  L  R  T  R  N  R  N  L  R  D  L  L  P  D  E  A
                                   *200

CACAGCTGGCGCGCTACGGCCTGCCGGTCTGGCGCACGGAAGAGGACGTGGCAGCGGCCC  660
 Q  L  A  R  Y  G  L  P  V  W  R  T  E  E  D  V  A  A  A  L

TGGGCGTCTCGGTGGGCGTGCTCCGCCACTACAGCATCCACCGCCCGCGCGAGCGGGTGC  720
 G  V  S  V  G  V  L  R  H  Y  S  I  H  R  P  R  E  R  V  R

GGCACTACGTGACCTTCGCCGTGCCCAAGCGCTCCGGAGGCGTCCGGCTGCTGCATGCGC  780
 H  Y  V  T  F  A  V  P  K  R  S  G  G  V  R  L  L  H  A  P
    *150

CCAAGCGGCGCCTGAAGGCCCTGCAACGCCGGATGCTGGCGCTCCTGGTGTCGAAGCTCC  840
 K  R  R  L  K  A  L  Q  R  R  M  L  A  L  L  V  S  K  L  P

CCGTGAGTCCACAGGCCCATGGCTTCGTGCCCGGCCGCTCCATCAAGACGGGCGCCGCGC  900
 V  S  P  Q  A  H  G  F  V  P  G  R  S  I  K  T  G  A  A  P
                                  *200

CGCACGTGGGCCGGCGGGTGGTCCTGAAGCTGGACCTGAAGGACTTCTTCCCCTCCGTCA  960
 H  V  G  R  R  V  V  L  K  L  D  L  K  D  F  F  P  S  V  T

CCTTCGCGCGGGTGCGAGGGCTGCTCATCGCCCTGGGCTACGGCTATCCCGTGGCGGCCA 1020
 F  A  R  V  R  G  L  L  I  A  L  G  Y  G  Y  P  V  A  A  T

CGCTCGCGGTGCTGATGACGGAGTCCGAGCGCCAGCCCGTGGAGCTGGAGGGCATCCTCT 1080
 L  A  V  L  M  T  E  S  E  R  Q  P  V  E  L  E  G  I  L  F
    *250

TCCACGTTCCCGTGGGCCCACGCGTCTGCGTGCAGGGCGCCCCCACGAGCCCCGCCCTGT 1140
 H  V  P  V  G  P  R  V  C  V  Q  G  A  P  T  S  P  A  L  C

GCAACGCGGTGCTGCTGCGACTGGACCGGCGGCTGGCGGGACTGGCGCGTCGGTACGGCT 1200
 N  A  V  L  L  R  L  D  R  R  L  A  G  L  A  R  R  Y  G  Y
                                    *300

ACACGTACACGCGCTACGCGGATGACCTCACCTTCTCCGGCGACGACGTCACGGCGCTGG 1260
 T  Y  T  R  Y  A  D  D  L  T  F  S  G  D  D  V  T  A  L  E

AGCGAGTCCGCGCGCTGGCCCCGCGGTACGTGCAGCAGGAAGGCTTCGAGCTCAACCGCG 1320
 R  V  R  A  L  A  A  R  Y  V  Q  E  E  G  F  E  V  N  R  E

AGAAGACCCGCGTGCAGCGCCGGGGCGGTGCCCAGCGCGTCACTGGCGTCACCGTGAATA 1380
 K  T  R  V  Q  R  R  G  G  A  Q  R  V  T  G  V  T  V  N  T
    *350

CGACGCTGGGCTTGTCACGCGAGGAGCGGCCGCGGCTCCGGGCGATGCTGCACCAGGAGG 1440
 T  L  G  L  S  R  E  E  R  P  R  L  R  A  M  L  H  Q  E  A

CGCGGTCGGAGGACGTCGAGGCACACCGCGCGCACCTCGACGGCCTCCTGGCCTACGTGA 1500
 R  S  E  D  V  E  A  H  R  A  H  L  D  G  L  L  A  Y  V  K
                                   *400

AGATGCTCAACCCGGAGCAGGCGGAGCGGCTCGCTCGCCGGCGCAAGCCGCGCGGACGT 1560
 M  L  N  P  E  Q  A  E  R  L  A  R  R  R  K  P  R  G  T  *

GAGCGAGGGCTCAGCTCCGGATGGGCCAGGGCCTGTCACGCGTCCCGGCCTCCCAGTTGT 1620

CATGGCGGCCGTCCCAGTAC
```

FIGURE 15

```
CC ACT TCC GGC GCT CGG GCT GCG CGA GGG CCC GTG CGA GCA CAT GAT GGC GCT GCG GCT     60
GT CCA GGT CCG GCA CCG CGC CGA GCA GGA AGC ACT GCG TCA GAC CCC CGC GGG CCG CCA    120
CT CAT CCG CGC GGA GAC GCG CTC CTA CGT GCG GCG CGA GCC CTC CGG CCA GGA GCA GGT    180
TA CCG CGT CTC ATT GGA TGG GAA AGT GGT GGC GGT GGA GTG GGG CCC CCG CCA GGG GGA    240
TC CCG CCG GCA GAA GCT CTG GTT CGA CAC GGA CGC CGA GGC GCG CAC CGC CTA CTT CAC    300
CG CCT GGA GTC CTT GGC CGC GGA GGG ATA TAT CGA TGC GGC TGC TTC AAT GAT GTA GAA    360
AC GCA AGC CAC GGG GCC GCG GGC GCG CGG CGG AAA GGC AGG TGC GAC GGA ACG ACA GAC    420
    #2          RNA⇨                                          ↓
CT CGT GCG AGC GAC CGA GAG AGG TCC CAA GCC ATC AGC CTC AGC GCC TCG AGC GCG AGA    480
GA GCA CGC TCG CTG GCT CTC TCC AGG GTT CGG TAG TCG GAG TCG CGG AGC TCG CGC TCT

CG GCG TTG CGC CGC TCT GGT TGA ATT GCA GGA CAC TCT CCG CAA GGT AGC CTG TTC TTG    540
GC CGC AAC GCG GCG AGA CCA ACT TAA CGT CCT GTG AGA GGC GTT CCA TCG GAC AAG AAC

T CTC TTC CCT CCG GTG AGT ACC TCT CCG GCC GGG GAG CTG AAC CAA CGA CGC AAC CGC     600
A GAG AAG GGA GGC CAC TCA TGG AGA GGC CGG CCC CTC GAC TTG GTT GCT GCG TTG GCG

T TTC CCC GGC CGG AGA GGT ACT CAC CGG AGG GGA GAG CCG GTG AGG CTA CCG TGC CCC     660
A AAG GGG CCG GCC TCT CCA TGA GTG GCC TCC CCT CTC GGC CAC TCC GAT GGC ACG GGG
                                #1
G TGA GAA GGT GGT GCC TTC GGG CCT CCC TCG ACC GCT CGC GCT CCG TCG CCC TGC CCT     720
C ACT CTT CCA CCA CGG AAG CCC GGA GGG AGC TGG CGA GCG CGA GGC AGC GGG ACG GGA
       ⇦DNA
CC TCG CCC CCC CCA CCT TGC TCA CCG GCG CCA GGA GCC GTC ATG ACC GCC AAG CTG GAG    780
                                                        M   T   A   K   L   E

CA CAC GTC CCC GCC GCG CCC CCC GTC TCC GCC GAG GCG CCC GCC CCC ACC CGT CCC GAT    840
S   H   V   P   A   A   P   P   V   S   A   E   A   P   A   P   T   R   P   D

CC GCG AAG CAG GAG GCC CGC CGC GCC CAC CAC GAG GCG CTG CGC CTG CGG TGG AAG GCC    900
A   A   K   Q   E   A   R   R   A   H   H   E   A   L   R   L   R   W   K   A

TC GAA GAG GCG GGC GGC ACG GAC GCC TGG GTG CGG CAG CAG CTG GTG GCC AAG GGC GTC    960
I   E   E   A   G   G   T   D   A   W   V   R   Q   Q   L   V   A   K   G   V
           *50
CG GCG GAA GAG GTG GAC TTC GAG TCG CTC AGC GAC AAG CAG AAG GCG GCC TGG AAG GAG   1020
A   A   E   E   V   D   F   E   S   L   S   D   K   Q   K   A   A   W   K   E

AG AAG AAG GCC GAG GCC ACC GAG CGG CGC GCG CAG AAG CGC CTG GCG TGG GAG GCC TGG   1080
K   K   K   A   E   A   T   E   R   R   A   Q   K   R   L   A   W   E   A   W
                                                   *100
AG GCC ACG CAC ATC CAC CAC CTG GGC GTG GGG GTG CAC TGG GAC GAG GCC GGA GGG CCG   1140
K   A   T   H   I   H   H   L   G   V   G   V   H   W   D   E   A   G   G   P

AC AAG TTC GAC GTG GCC GGG CGC GAG GAG CGG GCC AAG GCC AAC GGC TTG CCG GAG GGG   1200
D   K   F   D   V   A   G   R   E   E   R   A   K   A   N   G   L   P   E   G

TG GAC TCG GTC GAG GCG CTG GCC AAA GCG CTG GGC ATC TCC GTG TCG CGC CTG CGC TGG   1260
L   D   S   V   E   A   L   A   K   A   L   G   I   S   V   S   R   L   R   W
           *150
TC TCC TTC CAC CGC GAG GTG GAC ACG GGC ACG CAC TAC CAG ACG TGG GAG ATT CCG AAG   1320
F   S   F   H   R   E   V   D   T   G   T   H   Y   Q   T   W   E   I   P   K

GG GAC GGC GGC AAG CGG ACG CTC ACC GCG CCG AAG CGG GAG CTC AAG GCC GTG CAG CGC   1380
R   D   G   G   K   R   T   L   T   A   P   K   R   E   L   K   A   V   Q   R
                                                   *200
GG GTG CTC GCG AAC GTG GTG GAG CGG CTG CCG GTG CAC GGG GCC GCG CAC GGC TTC GTG   1440
W   V   L   A   N   V   V   E   R   L   P   V   H   G   A   A   H   G   F   V
```

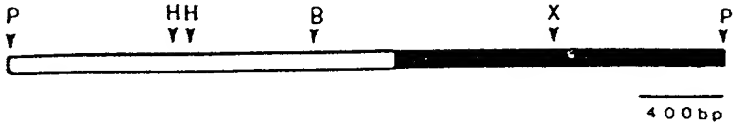
```
GCG GGG CGC TCC ATC CTC ACC AAC GCG CTG GCC CAC CAG GGC GCG GAC GTG GTG GTG AAG   1500
 A   G   R   S   I   L   T   N   A   L   A   H   Q   G   A   D   V   V   V   K

GTG GAC ATG AAG GAC TTC TTC CCT TCC GTG ACG TGG CCC CGG GTC AAG GGA CTG CTG CGC   1560
 V   D   M   K   D   F   F   P   S   V   T   W   P   R   V   K   G   L   L   R
            *250
AAG GGA GGA CTC CCG GAG AAC CTG GCG ACG CTC CTG GCG CTG CTC TCC ACC GAG GCC CCG   1620
 K   G   G   L   P   E   N   L   A   T   L   L   A   L   L   S   T   E   A   P

CGC GAG GTG GTG CGG TTC CGG GGA GAG ACG CTG TAC GTG GCC AAG GGC CCT CGC GCG CTG   1680
 R   E   V   V   R   F   R   G   E   T   L   Y   V   A   K   G   P   R   A   L
                                                    *300
CCC CAG GGG GCC CCC ACC TCT CCG GCG CTG ACG AAC GCG CTG TGC CTG CGG CTG GAC AAG   1740
 P   Q   G   A   P   T   S   P   A   L   T   N   A   L   C   L   R   L   D   K

CGG CTC TCG GCG CTG TCG AAG CGG CTG GGC TTC ACG TAC ACG CGC TAT GCG GAT GAC CTG   1800
 R   L   S   A   L   S   K   R   L   G   F   T   Y   T   R   Y   A   D   D   L

ACG TTC TCC TGG CGG CGG GCG AAG AAG TCC CGG CAG AAG GAA CTC CCC CTG GCG GAT GCG   1860
 T   F   S   W   R   R   A   K   K   S   R   Q   K   E   L   P   L   A   D   A
            *350
CCG GTG GCG CTG CTC CTG GCG CGG GTG AAG GGT GTG CTG GAG GCC GAG GGT TTC ACG CTG   1920
 P   V   A   L   L   L   A   R   V   K   G   V   L   E   A   E   G   F   T   L

CAC CCG GAC AAG ACG CGG GTG CAG CGC AAG GGC AGC CGG CAG CGG GTG ACG GGG CTC GTG   1980
 H   P   D   K   T   R   V   Q   R   K   G   S   R   Q   R   V   T   G   L   V
                                                    *400
GTG AAC GAG GCC CCC GAG GGC GTT CCG GGT GCC CGG GTG CCC CGC GAT GTG GTG CGG CGG   2040
 V   N   E   A   P   E   G   V   P   G   A   R   V   P   R   D   V   V   R   R

CTG CGC GCG GCG ATC CAC AAC CGG GAG CAG GGC AAG CCC GGC CCC ACC GGG GAG ACG CTG   2100
 L   R   A   A   I   H   N   R   E   Q   G   K   P   G   P   T   G   E   T   L

GAG CAG CTC AAG GGG CTC GCG GCC TTC CTT CAC ATG ACG GAC GCG GAG AAG GGC CGC GCC   2160
 E   Q   L   K   G   L   A   A   F   L   H   M   T   D   A   E   K   G   R   A
            *450
TTC CTG CGA CGG CTG GAG GCC CTC GAG AAG CGC CAG ACC GCC TGA CCC TCA CTG GTC GTC   2220
 F   L   R   R   L   E   A   L   E   K   R   Q   T   A   -

CGG GGC ATC GCA GCG GGC GCC GGG ACG GAC CGT CAC CCC CCA GAT CTC CAT GCC ATG CTG   2280
                                                                  ↓
GGG ATT CTG GGC GGT GAA GAA GAC TTC CCA GCC GAG ACG GAC GAA GCC CTG CGG ATC CGA   2340
TGA CTC CTC GCC CGG GGC GAT CTC CCG GAG GGG CAC CGT TCC GAC GTC CGT GCC ATT GCT   2400
CAC CCA GGG CTC CCG GCC CCA GCC TTG GGT GTC CGC CGA GAA GAA GAG CAG CCC GGA GAT   2460
GGC CGT CAG GTT CTC CGG CGA CGC ATC CTC GGG GCC CGG CGC CAA ATC CTT CAG CAG CAG   2520
GGT GCC CTT GGC GGT GCC ATC GCT GGA CCA CAG CTC CCG GCC GTG GAG GCT GTC ACT CGC   2580
GGC GAA GTA GAG CAT CCC ATT CAG CGC CTT GAT GGC GCT GGG CGC CGA GCT GTC CGG ACC   2640
CGG CCA GAT GTC CTT CAC CCG GAC CGT GCC ATG CGA CGT GCC ATC GCT GAC CCA CAG CTC   2700
CTC GCC CTC GGG CTG GCC CCA GAA CTC GGG CTC GCC TCC CCC GGC GCT GAA GAA GAT CTT   2760
CCC CCC GAG CGC CGT GAG ATC ATG CGG ATA GAG GCC GGG GAA GAA GCG CAG CTG CTC GGA   2820
GAC GGT GCC TCT GGA GCA CCA CAG GCT GGC CTC GCC TTC GTC ATT GTC GAG CAG GAA GAA   2880
GAG CAC CGA GTC CGC CGC GGT GAA CGC GGA GAG GAA GTT GTC CTC GGG GCC CGT GAA GAC   2940
AGA CGT GGT GCT GGA CAG CCC CAG GCT GCG CCA GAT GAA CAC CTC GTC ATT GAC GTT GGC   3000
CAC GAA GAA GAG CGC ATC GCC GAC CCG GGT GAG CCG GCG CGG GCT GGA GCT GCC GGG CAC   3060
```

FIGURE 16

```
TTTCGAGAAG CGCCATACCA AACAGGGGAT ACAGACCAAC CTGACGCTGA AAGAGGAAAG CTACGGCGAC TGGCTGCCGA AGTGCCACGA 9990
 F  E  K  R  H  T  K  Q  G  I  Q  T  N  L  T  L  K  E  E  S  Y  G  D  W  L  P  K  C  D  D
CCCCGCAGCA ACATAACCTC ACTCAGACCG GCAACAGCCG GTCTTTTCCT TTCTGGCCAT TGCCACAAGG TGAACAATCC ACTGTTCACC 10080
 P  A  A  T  *
CTTCACCGTT TATTCACCCT TTATCACTAT GAAATTATTA ATAAAAAACC AGAGGTGAAC AGTGTGAACA GTAAAACCTG AAAAAACTTT 10170
TTATCACCCC GCGCATCGCC CGACTGGACA GATCCAGAAC GAGCAAAAAT CACAAAGGTG ACGAGTCGAC TGTTCACTCT TCACCAACTC 10260
 crf
ATCACCACCT AACCACATGA TATAAAATGA TAAATAATCG AGGTGAACAG TTAAATGCAA AAAAACTTTT TCTCAGCTCT TGGATAAAAG 10350
                                                             RNA⇦    #2
AAAATTAATT CACATCAATA GCTTTCCTCT TGAATCCTCT TGAGGTTTAT GAGAGCGTAA CAGAGCCAAA CCTAGCATTT TATGGGTTAA 10440
TTTTAATTAA GTGTAGTTAT CGAAAGGAGA ACTTAGGAGA ACTCCAAATA CTCTCGCATT GTCTCGGTTT GGATCGTAAA ATACCCAATT

TAGCCCATCG CGCATGAGTC ATGGTTTCGC CTAGTATTTT AGCTATGCCC GTCGTTCAGT TCGCTGAGCG GCGGCTGGGG GCCACCGATC 10530
ATCGGGTAGC GCGTACTCAG TACCAAAGCG GATCATAAAA TCGATACGGG CAGCAAGTCA AGCGACTCGC CGCCGACCCC CGGTGGCTAG
                                #1
AGCGAACTGA TCGACGTGCT CAAGTAGGTT TGGCTCTTTT AGTCCTCTAC CATCAAGGTG CATAAGGATA TTCTCGATGC TGACTCAGCT 10620
TCGCTTGACT AGCTGCACGA GTTCATCCAA ACCGAGAAAA TCAGGAGATG GTAGTTCCAC GTATTCCTAT AAGAGCTACG ACTGAGTCGA
                          ⇦DNA                                                orf316  M  L  T  Q  L
AAAAAAAAAT GGTACTGAGG TATCTAGAGC AACCGCGTTA TTTTCATCAT TCGTTGAAAA GAACAAAGTA AAATGTCCTG GTAATGTAAA 10710
 K  K  N  G  T  E  V  S  R  A  T  A  L  F  S  S  F  V  E  K  N  K  V  K  C  P  G  N  V  K
AAAATTCGTC TTTCTGTGTG GTGCTAACAA AAACAATGGA GAACCATCAG CAAGACGATT GGAATTAATA AATTTTTCTG AAAGGTATTT 10800
 K  F  V  F  L  C  G  A  N  K  N  N  G  E  P  S  A  R  R  L  E  L  I  N  F  S  E  R  Y  L
                                                                                           IV
GAATAACTGT CACTTTTTTC TTGCTGAACT AGTTTTCAAA GAATTAAGCA CCGATGAAGA ATCATTATCT GATAATTTAT TAGATATCGA 10890
 N  N  C  H  F  F  L  A  E  L  V  F  K  E  L  S  T  D  E  E  S  L  S  D  N  L  L  D  I  E
AGCTGACTTA TCTAAATTAG CTGATCATAT TATCATTGTT TTAGAAAGTT ATTCATCTTT CACGGAACTT GGTGCATTCG CATACAGCAA 10980
 A  D  L  S  K  L  A  D  H  I  I  I  V  L  E  S  Y  S  S  F  T  E  L  G  A  F  A  Y  S  K
GCAATTACGC AAGAAATTAA TAATAGTTAA CAATACAAAA TTTATAAATG AGAAATCATT TATAAATATG GGACCAATAA AGGCTATTAC 11070
 Q  L  R  K  K  L  I  I  V  N  N  T  K  F  I  N  E  K  S  F  I  N  N  G  P  I  K  A  I  T
TCAGCAATCA CAACAATCTG GTCATTTCTT ACATTATAAA ATGACAGAAG GTATTGAAAG TATAGAGCGC TCTGATGGGA TTGGCGAAAT 11160
 Q  Q  S  Q  Q  S  G  H  F  L  H  Y  K  M  T  E  G  I  E  S  I  E  R  S  D  G  I  G  E  I
ATTCGACCCC CTATATGATA TTCTTTCTAA GAACGACAGA GCAATTTCAA GAACTTTAAA AAAAGAAGAG TTAGATCCTT CCAGTAACTT 11250
 F  D  P  L  Y  D  I  L  S  K  N  D  R  A  I  S  R  T  L  K  K  E  E  L  D  P  S  S  N  F
CAATAAAGAC TCAGTACGAT TTATTCATGA CGTAATTTTT GTATGTGGTC CTTTGCAACT TAATGAACTC ATCGAAATAA TCACAAAAAT 11340
 N  K  D  S  V  R  F  I  H  D  V  I  F  V  C  G  P  L  Q  L  N  E  L  I  E  I  I  T  K  I
                                                 II
ATTTGGCACA GAAAGCCATT ACAAAAAAAA TCTTCTAAAG CACCTTGGTA TTCTAATAGC TATTAGAATA ATATCATGCA CAAATGGGAT 11430
 F  G  T  E  S  H  Y  K  K  N  L  L  K  H  L  G  I  L  I  A  I  R  I  I  S  C  T  N  G  I
TTATTATTCT TTGTATAAAG AATATTATTT TAAATATGAC TTTGACATTG ACAACATATC ATCAATGTTT AAAGTTTTTT TCCTCAAGAA 11520
 Y  Y  S  L  Y  K  E  Y  Y  F  K  Y  D  F  D  I  D  N  I  S  S  M  F  K  V  F  F  L  K  N
CAAGCCAGAA AGGATGAGGG TATATGAGAA TATATAGCCT AATTGATTCT CAGACATTGA TGACTAAGGG ATTTGCTTCT GAAGTAATGC 11610
 K  P  E  R  N  R  V  Y  E  N  I  *
                     RT  M  R  I  Y  S  L  I  D  S  Q  T  L  M  T  K  G  F  A  S  E  V  M
GATCACCTGA GCCGCCAAAA AAATGGGATA TAGCTAAGAA AAAAGGAGGT ATGAGAACAA TTTATCACCC GTCATCAAAA GTTAAATTAA 11700
 R  S  P  E  P  P  K  K  W  D  I  A  K  K  K  G  G  M  R  T  I  Y  H  P  S  S  K  V  K  L
TTCAATATTG GTTAATGAAT AATGTTTTTT CGAAGCTCCC AATGCATAAT GCTGCATATG CATTTGTTAA AAACCGATCA ATAAAAAGCA 11790
 I  Q  Y  W  L  M  N  N  V  F  S  K  L  P  M  H  N  A  A  Y  A  F  V  K  N  R  S  I  K  S
ATGCTTTATT ACATGCCGAA TCAAAGAATA AGTATTATGT GAAAATAGAT CTCAAAGATT TTTTCCCTTC AATAAAATTT ACTGATTTTG 11880
 N  A  L  L  H  A  E  S  K  N  K  Y  Y  V  K  I  D  L  K  D  F  F  P  S  I  K  F  T  D  F
AGTACGCATT CACTCGTTAT CGAGATCGCA TTGAATTTAC TACAGAATAT GATAAGGAGT TACTACAACT TATAAAAACG ATCTGCTTTA 11970
 E  Y  A  F  T  R  Y  R  D  R  I  E  F  T  T  E  Y  D  K  E  L  L  Q  L  I  K  T  I  C  F
TATCAGATAG CACTCTCCCT ATCGGGTTTC CTACATCTCC ATTAATTGCA AACTTTGTGG CAAGAGAACT TGATGAAAAA CTGACGCAAA 12060
 I  S  D  S  T  L  P  I  G  F  P  T  S  P  L  I  A  N  F  V  A  R  E  L  D  E  K  L  T  Q
AACTAAATGC AATTGATAAA CTTAATGCCA CTTATACACG ATATGCTGAT GATATTATTG TCTCTACAAA TATGAAAGGG GCTAGCAAAT 12150
 K  L  N  A  I  D  K  L  N  A  T  Y  T  R  Y  A  D  D  I  I  V  S  T  N  M  K  G  A  S  K
TAATTCTGGA TTGTTTTAAA AGAACAATGA AAGAGATTGG TCCAGACTTT AAAATTAACA TTAAAAAATT TAAGATTTGT AGTGCTTCGG 12240
 L  I  L  D  C  F  K  R  T  M  K  E  I  G  P  D  F  K  I  N  I  K  K  F  K  I  C  S  A  S
GAGGAAGTAT AGTAGTTACC GGATTGAAAG TTTGCCACGA TTTTCATATT ACATTACATA GATCAATGAA AGATAAAATA AGATTGCATC 12330
 G  G  S  I  V  V  T  G  L  K  V  C  H  D  F  H  I  T  L  H  R  S  M  K  D  K  I  R  L  H
TTTCTCTTTT ATCAAAGGGC ATATTAAAAG ATGAAGATCA TAATAAACTT TCTGGTTATA TTGCTTATGC AAAAGATATA GACCCTCATT 12420
 L  S  L  L  S  K  G  I  L  K  D  E  D  H  N  K  L  S  G  Y  I  A  Y  A  K  D  I  D  P  H
TTTATACAAA ACTGAACAGA AAATATTTTC AAGAAATAAA ATGGATTCAG AATCTCCACA ACAAAGTTGA ATAAACTTTA TATTTTGGAT 12510
 F  Y  T  K  L  N  R  K  Y  F  Q  E  I  K  M  I  Q  N  L  H  N  K  V  E  *
GCACCCCAAT AACTTCATTG ATTAAATTGG GAACAATATA GGCTTTTCAG GATGACCTAC ACTCTAGAGA ATGTGTATAC AAAAGTGTAT 12600
AAGTTATTTT CAAACCTATA TAAAATACAG CAAAATCAAT GCATTGGCGG CATTTTACCA CTCCTGTGAT CTTCCGCCAA AATGCCTG  12688
```

FIGURE 17

(A)



```
-371  TGGCATCTATTAAGAAGGTTAGGAAAGAAAATAAAGTATCAAAAGATATTGGAAATATAT

-311  TATACGCAGAGCGTTTCTATTGCCTTGTATCTATTTACTGGATAGTGTCAACTACCGCAC

-251  ACTGTGTGAACTAGCTTTTAAAGCGATAAAGCAAGATGATGTTTTATCTAAAATTATTGT

-191  TAGATCCGTTGTTTCTCGTCTAATAAATGAACGAAAAATACTTCAAATGACTGATGGTTA

-131  TCAGGTCACTGCTTTGGGGGCTAGCTATGTTAGGAGCGTCTTTGATAGAAAGACACTTGA

 -71  CCGATTGCGGCTTGAGATTATGAATTTTGAAAACCGTAGAAAATCAACATTTAACTATGA
                +1
                |-|-------------------msdRNA-----------------------
 -11  TAAGATTCCGTATGCGCACCCTTAGCGAGAGGTTTATCATTAAGGTCAACCTCTGGATGT
                IR ============
      ---------------------------------->
  49  TGTTTCGGCATCCTGCATTGAATCTGAGTTACTGTCTGTTTTCCTTGTTGGAACGGAGAG
                               <---------------------------------------

 109  CATCGCCTGATGCTCTCCGAGCCAACCAGGAAACCCGTTTTTCTGACGTAAGGGTGCGC
      ----------------------msDNA----------------------| ===========

 169  AACTTTCATGAAATCCGCTGAATATTTGAACACTTTTAGATTGAGAAATCTCGGCCTACC
      = IR    MetLysSerAlaGluTyrLeuAsnThrPheArgLeuArgAsnLeuGlyLeuPr

 229  TGTCATGAACAATTTGCATGACATGTCTAAGGCGACTCGCATATCTGTTGAAACACTTCG
      oValMetAsnAsnLeuHisAspMetSerLysAlaThrArgIleSerValGluThrLeuAr

 289  GTTGTTAATCTATACAGCTGATTTTCGCTATAGGATCTACACTGTAGAAAAGAAAGGCCC
      gLeuLeuIleTyrThrAlaAspPheArgTyrArgIleTyrThrValGluLysLysGlyPr

 349  AGAGAAGAGAATGAGAACCATTTACCAACCTTCTCGAGAACTTAAAGCCTTACAAGGATG
      oGluLysArgMetArgThrIleTyrGlnProSerArgGluLeuLysAlaLeuGlnGlyTr

 409  GGTTCTACGTAACATTTTAGATAAACTGTCGTCATCTCCTTTTTCTATTGGATTTGAAAA
      pValLeuArgAsnIleLeuAspLysLeuSerSerSerProPheSerIleGlyPheGluLy

 469  GCACCAATCTATTTTGAATAATGCTACCCCGCATATTGGGGCAAACTTTATACTGAATAT
      sHisGlnSerIleLeuAsnAsnAlaThrProHisIleGlyAlaAsnPheIleLeuAsnIl

 529  TGATTTGGAGGATTTTTTCCCAAGTTTAACTGCTAACAAAGTTTTTGGAGTGTTCCATTC
      eAspLeuGluAspPhePheProSerLeuThrAlaAsnLysValPheGlyValPheHisSe

 589  TCTTGGTTATAATCGACTAATATCTTCAGTTTTGACAAAAATATGTTGTTATAAAAATCT
      rLeuGlyTyrAsnArgLeuIleSerSerValLeuThrLysIleCysCysTyrLysAsnLe

 649  GCTACCACAAGGTGCTCCATCATCACCTAAATTAGCTAATCTAATATGTTCTAAACTTGA
      uLeuProGlnGlyAlaProSerSerProLysLeuAlaAsnLeuIleCysSerLysLeuAs

 709  TTATCGTATTCAGGGTTATGCAGGTAGTCGGGGCTTGATATATACGAGATATGCCGATGA
      pTyrArgIleGlnGlyTyrAlaGlySerArgGlyLeuIleTyrThrArgTyrAlaAspAs

 769  TCTCACCTTATCTGCACAGTCTATGAAAAAGGTTGTTAAAGCACGTGATTTTTTATTTTC
      pLeuThrLeuSerAlaGlnSerMetLysLysValValLysAlaArgAspPheLeuPheSe

 829  TATAATCCCAAGTGAAGGATTGGTTATTAACTCAAAAAAAACTTGTATTAGTGGGCCTCG
      rIleIleProSerGluGlyLeuValIleAsnSerLysLysThrCysIleSerGlyProAr

 889  TAGTCAGAGGAAAGTTACAGGTTTAGTTATTTCACAAGAGAAAGTTGGGATAGGTAGAGA
      gSerGlnArgLysValThrGlyLeuValIleSerGlnGluLysValGlyIleGlyArgGl

 949  AAAATATAAAGAAATTAGAGCAAAGATACATCATATATTTTGCGGTAAGTCTTCTGAGAT
      uLysTyrLysGluIleArgAlaLysIleHisHisIlePheCysGlyLysSerSerGluIl

1009  AGAACACGTTAGGGGATGGTTGTCATTTATTTTAAGTGTGGATTCAAAAAGCCATAGGAG
      eGluHisValArgGlyTrpLeuSerPheIleLeuSerValAspSerLysSerHisArgAr

1069  ATTAATAACTTATATTAGCAAATTAGAAAAAAAATATGGAAAGAACCCTTTAAATAAAGC
      gLeuIleThrTyrIleSerLysLeuGluLysLysTyrGlyLysAsnProLeuAsnLysAl

1129  GAAGACCTAATGGTCTTCGTTTTAAAACTAAAGCTCATAGGTTGAAAAATTGAGCACTTC
      aLysThr

1189  TTCGTCCAACCAGTTATTTAGTTCCTGCAATCGTTTCTGCAG
```

FIGURE 18

```
     Oligo 2337
tcaccctgaaagacctgattgcttacctggaagagaagccggaaatggcggaacatctgg   60

cggcggttaaggcctatcgcgaagagttcggcgtttaaaAATATGCGCTGTGCAGGGTTT   120
                                          RNA      a2
TTGCTGTGCGCAGCGTGATGCGCTTCAAGATATCGTGTTAATCTGCTTTCGCCAGCAGTG   180
AACGACACGCGTCGCACTACGCGAAGTTCTATAGCACAATTAGACGAAAGCGGTCGTCAC

GCAATAGCGTTTCCGGCCTTTTGTGCCGGGAGGGTCGGCGAGTCGCTGACTTAACGCCAG   240
CGTTATCGCAAAGGCCGGAAAACACGGCCCTCCCAGCCGCTCAGCGACTGAATTGCGGTC

TAGTATGTCCATATACCCAAAGTCGCTTCATTGTACCTGAGTACGCTTCGCGTACGTCGC   300
ATCATACAGGTATATGGGTTTCAGCGAAGTAACATGGACTCATGCGAAGCGCATGCAGCG
                                                         a1
GCTGACGCGCTCAGTACAGTTACGCGCCTTCGGGATGGTTTAATGGTATTGCCGCTGTTG   360
CGACTGCGCGAGTCATGTCAATGCGCGGAAGCCCTACCAAATTACCATAACGGCGACAAC
                                             DNA
GCGCCTCTTTTGGCCGCCGTGATGTGGAGAGTGGAATGGATGCTACCCGGACAACCCTTC   420
                                             M  D  A  T  R  T  T  L  L

TGGCGCTCGATTTGTTCGGCTCGCCGGGCTGGAGCGCCGATAAAGAAATACAGCGACTGC   480
 A  L  D  L  F  G  S  P  G  W  S  A  D  K  E  I  Q  R  L  H

ATGCGCTCAGTAATCATGCCGGACGCCATTACCGACGCATTATTCTTTCTAAACGCCACG   540
 A  L  S  N  H  A  G  R  H  Y  R  R  I  I  L  S  K  R  H  G

GTGGTCAGCGGCTGGTGTTAGCCCCTGATTACTTGCTCAAAACCGTACAGCGCAACATTC   600
 G  Q  R  L  V  L  A  P  D  Y  L  L  K  T  V  Q  R  N  I  L

TTAAGAACGTCCTTTCACAATTTCCGCTTTCCCCTTTTGCTACAGCCTACCGACCAGGTT   660
 K  N  V  L  S  Q  F  P  L  S  P  F  A  T  A  Y  R  P  G  C

GCCCAATCGTCAGCAACGCGCAGCCACACTGCCAACAGCCGCAGATCCTGAAACTCGATA   720
 P  I  V  S  N  A  Q  P  H  C  Q  Q  P  Q  I  L  K  L  D  I

TCGAAAACTTTTTCGATAGCATTAGCTGGTTACAGGTCTGGCGTGTGTTTCGCCAGGCCC   780
 E  N  F  F  D  S  I  S  W  L  Q  V  W  R  V  F  R  Q  A  Q

AGTTGCCACGTAATGTGGTAACCATGCTGACCTGGATTTGTTGTTATAACGACGCGTTAC   840
 L  P  R  N  V  V  T  M  L  T  W  I  C  C  Y  N  D  A  L  P

CGCAGGGGGCACCAACTTCGCCAGCCATTTCCAATCTTGTGATGCGCCGTTTTGATGAAC   900
 Q  G  A  P  T  S  P  A  I  S  N  L  V  M  R  R  F  D  E  R

GCATAGGGGAATGGTGTCAGGCTCGGGGAATTACCTACACCCGCTACTGCGATGACATGA   960
 I  G  E  W  C  Q  A  R  G  I  T  Y  T  R [Y  C  D  D] M  T

CCTTTTCAGGTCACTTCAATGCCCGCCAGGTTAAAAATAAAGTGTGCGGATTGTTAGCGG   1020
 F  S  G  H  F  N  A  R  Q  V  K  N  K  V  C  G  L  L  A  E

AGCTGGGCCTGAGCCTCAATAAACGCAAAGGCTGCCTGATAGCTGCCTGTAAGCGCCAGC   1080
 L  G  L  S  L  N  K  R  K  G  C  L  I  A  A  C  K  R  Q  Q

AAGTAACCGGGATTGTTGTTAATCACAAGCCACAGCTTGCCCGTGAAGCGCGCCGGGCGC   1140
 V  T  G  I  V  V  N  H  K  P  Q  L  A  R  E  A  R  R  A  L

TGCGTCAGGAGGTGCATTTGTGCCAAAAATATGGCGTTATTTCGCATCTTAGTCATCGTG   1200
 R  Q  E  V  H  L  C  Q  K  Y  G  V  I  S  H  L  S  H  R  G

GTGAACTTGATCCTTCTGGCGATCTCCACGCACAGGCAACGGCGTATCTTTATGCTTTGC   1260
 E  L  D  P  S  G ·D  L  H  A  Q  A  T  A  Y  L  Y  A  L  Q

AGGGAAGAATAAACTGGTTATTGCAAATCAACCCTGAGGATGAGGCCTTTCAACAGGCGA   1320
 G  R  I  N  W  L  L  Q  I  N  P  E  D  E  A  F  Q  Q  A  R

GAGAGAGTGTAAAGCGAATGCTGGTTGCATGGTAAGAAAAGCGTCAGGCAGACGTTTCTG   1380
 E  S  V  K  R  M  L  V  A  W  *

CCTGACCGTTTAGGGGAGAattactgcaactgcgcggcaattagcggccagcgggcgtca   1440

aaatcatccgtcgggcggtatttaaactcgctgcggacaaaacgtgacagcataccttca   1500

cagaaggccaggatctggcttgccagcagggtttcatcgg                       1540
                             Oligo 2336
```

FIGURE 19

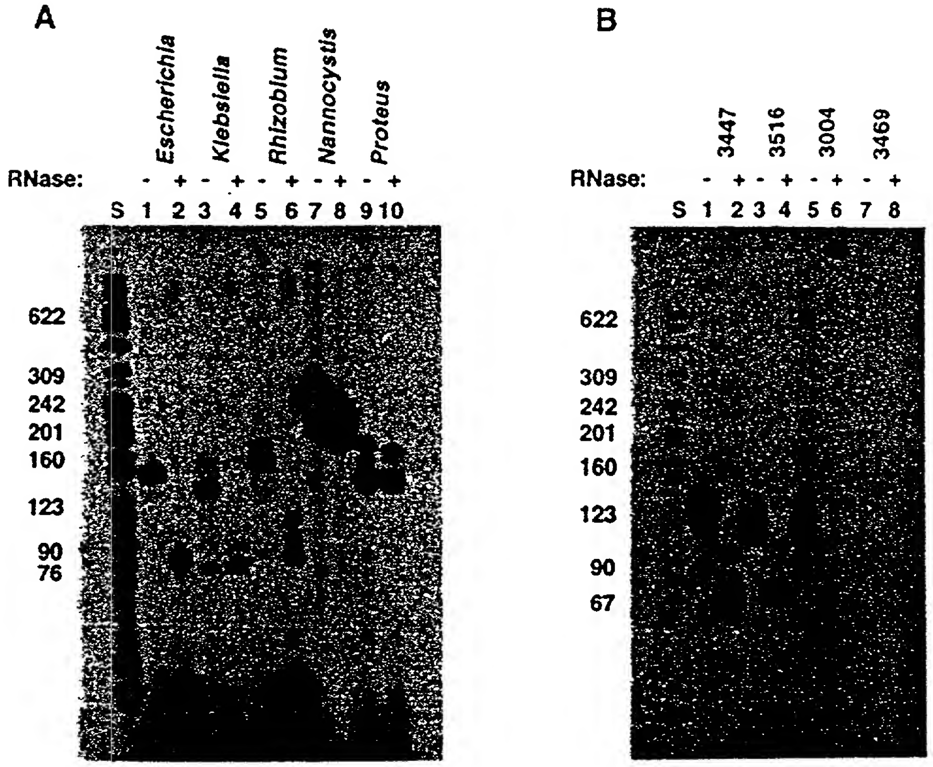

A
Escherichia
Klebsiella
Rhizobium
Nannocystis
Proteus
RNase:
- + - + - + - + - +
S 1 2 3 4 5 6 7 8 9 10
622
309
242
201
160
123
90
76
B
3447
3516
3004
3469
RNase:
- + - + - + - +
S 1 2 3 4 5 6 7 8
622
309
242
201
160
123
90
67

FIGURE 20

RHIZOBIAL ISOLATES

Strain (legume host genus)	USDA strain no.	Geographic source (date)	msDNA produced[b]
Rhizobium sp. (*Acacia*)	3002	Brazil (1959)	+
	3003	Africa (1950)	
	3325	Morocco (1974)	
	3838	? (1976)	+
Bradyrhizobium sp. (*Aeschynomene*)	3516	Florida (1972)	+
	4362		
Bradyrhizobium sp. (*Albizia*)	3004	Maryland (1952)	+
Bradyrhizobium sp. (*Apios*)	3240	Maryland (1939)	
Bradyrhizobium sp. (*Arachis*)	3339	Thailand (1979)	
	3341	Hawaii (1978)	
Rhizobium sp. (*Astragalus*)	3854	Alaska (1962)	
Rhizobium sp. (*Cajanus*)	3472		
Bradyrhizobium sp. (*Canavalia*)	3317	Brazil (1974)	
Rhizobium sp. (*Cicer*)	3378		
	3379	Mexico (1963)	
Bradyrhizobium sp. (*Coronilla*)	3165	Virginia (1935)	
	3167	? (1961)	
Bradyrhizobium sp. (*Crotalria*)	3384	Brazil (1967)	
Bradyrhizobium sp. (*Desmodium*)	3225	Ecuador (1948)	
Bradyrhizobium sp. (*Erythrina*)	3241		
	3242	Maryland (1939)	+
Rhizobium fredii	191	China (1979)	
Rhizobium leguminosarum	2370	Illinois (1933)	
	2429	Hawaii (1978)	
	2435	Holland (1955)	
	2480	Tennessee (1951)	
	2489		
Rhizobium sp. (*Lens*)	2426		
	3404	Colombia (1979)	
Rhizobium loti	3084	Maryland (1946)	
	3468	New Zealand (1961)	+
	3469		
	3471		
	3503		+
	3669	California (1968)	
Bradyrhizobium sp. (*Lotus*)	3074	Minnesota (1954)	
	3470	California (1916)	
Rhizobium sp. (*Lupinas*)	3040	Florida (1940)	
Bradyrhizobium sp. (*Lupinas*)	3045	Florida (1946)	
Bradyrhizobium sp. (*Macrotyloma*)	3451	Zimbabwe (1960)	
Rhizobium medicago	1097	North Dakota (1948)	
Rhizobium meliloti	1011	Maryland (1933)	
	1021a	North Dakota (1948)	
Rhizobium phaseoli	2667	Washington (1948)	
	2669		
	2674	Brazil (?)	
	2676	Colombia (1972)	
	3256	Illinois (1941)	
Rhizobium sp. (*Robinia*)	3436		
Bradyrhizobium sp. (*Stylosanthes*)	3441	Brazil (?)	
	3477	Colombia (1976)	
Rhizobium trifolii	2046	Virginia (1934)	
	2048	Illinois (1934)	+
	2063	Florida (1939)	
	2065	Alabama (1952)	+
	2116	South Carolina (1944)	
	2134	? (1974)	
	2145		
	2156	California (1920)	
Rhizobium sp. (*Trigonella*)	1177	Florida (1939)	
Rhizobium tropici	2744	Brazil (?)	
Bradyrhizobium sp. (*Vigna*)	3447	Thailand (1979)	+
	3456	Wisconsin (1966)	

[a] All strains are from the USDA Beltsville Rhizobium Culture Collection, provided by Peter van Berkum.
[b] As defined by detection of radiolabeled msDNA by the RT extension method.

FIGURE 21

SEQUENCE LISTING

Sub G'

(1) GENERAL INFORMATION:

(i) APPLICANT: Inouye, Sumiko
Hsu, Mei-Yin
Eagle, Susan
Inouye, Masayori

(ii) TITLE OF INVENTION: Prokaryotic Reverse Transcriptase

(iii) NUMBER OF SEQUENCES: 45

C1

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: Weiser & Associates
(B) STREET: 230 South Fifteenth Street, Suite 500
(C) CITY: Philadelphia
(D) STATE: Pennsylvania
(E) COUNTRY: U.S.A.
(F) ZIP: 19102

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.25

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER: US 08/269,118
(B) FILING DATE: 30-JUN-1994
(C) CLASSIFICATION:

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: Weiser, Gerard J.
(B) REGISTRATION NUMBER: 19,763
(C) REFERENCE/DOCKET NUMBER: 377.5888P

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 215-875-8383
(B) TELEFAX: 215-875-8394

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2176 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ix) FEATURE:
(A) NAME/KEY: CDS

```
        (B) LOCATION: 640..2094

    (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

TCATCCGCGC GGACACCCCC TCCTACGTGC CCCCCGACGC GGAGAGCGGC GTGGAGACGG      60

TGTACCGCGT TTCCCTGGAT GGTCACCTGG TGGCGGTGGA GTGGGGCCCG CGCACGGGCT     120

CGCCGCGTCA CCAGCGGCTC TGGTTCGACT CGGATGCGGA AGCCCCCGGA GCCTACTTCG     180

CGCGCCTCGA GAAGTTGGCG GCTGACGGCT ACATCGACGC GGCCTCGGCA TTGGTCTAAA     240

CCCTTCAACC ACGGCTCGGC CGCCACGCGC GGCCGGCAGG ACAGGTGCGA CGAACAGACG     300

ACGACGTGCG CTTCACGCGC GAGCAGCCGA GAGAGGTCCG GAGTGCATCA GCCTGAGCGC     360

CTCGAGCGGC GGAGCGGCGT TGCGCCGCTC CGGTTGGAAT GCAGGACACT CTCCGCAAGG     420

TAGCCTGTTC TTGGCTCTCT CCCTCCTAGG CACTACGGCC AGGGTGGGTA GCGGAGCCAA     480

CGACGCCACC GCCGTTTACC CACCCCGGCC GTAGTGCCTA GGAGGGGAGA GCCGGTGAGG     540

CTACCGTGCC CCAGGTAAGA TGGTGGTGCT TTCCCGGCCT CCGTCGACTG CTCGCGCCAT     600

GTCCCGTCTT CCATCGCCGC GCCCGCCCAA GGTGCAGAC ATG ACC GCC AGG CTG       654
                                          Met Thr Ala Arg Leu
                                           1               5

GAC CCG TTC GTC CCC GCA GCT TCG CCG CAG GCC GTG CCC ACG CCC GAG       702
Asp Pro Phe Val Pro Ala Ala Ser Pro Gln Ala Val Pro Thr Pro Glu
                 10                  15                  20

CTC ACC GCT CCG TCG TCA GAC GCG GCC GCG AAG CGT GAA GCC CGC CGG       750
Leu Thr Ala Pro Ser Ser Asp Ala Ala Ala Lys Arg Glu Ala Arg Arg
             25                  30                  35

CTC GCG CAC GAA GCG TTG CTC GTC CGC GCG AAG GCC ATC GAC GAA GCG       798
Leu Ala His Glu Ala Leu Leu Val Arg Ala Lys Ala Ile Asp Glu Ala
         40                  45                  50

GGC GGC GCC GAC GAC TGG GTG CAG GCG CAG CTC GTC TCC AAG GGG CTC       846
Gly Gly Ala Asp Asp Trp Val Gln Ala Gln Leu Val Ser Lys Gly Leu
     55                  60                  65

GCG GTC GAG GAC CTG GAC TTC TCC AGC GCC TCC GAG AAG GAC AAG AAG       894
Ala Val Glu Asp Leu Asp Phe Ser Ser Ala Ser Glu Lys Asp Lys Lys
 70                  75                  80                  85

GCC TGG AAG GAG AAG AAG AAG GCC GAG GCC ACC GAG CGC CGC GCG CTG       942
Ala Trp Lys Glu Lys Lys Lys Ala Glu Ala Thr Glu Arg Arg Ala Leu
                 90                  95                 100

AAG CGT CAG GCG CAC GAG GCG TGG AAG GCC ACG CAC GTG GGC CAC CTG       990
```

C1

Sub G1 cont.

C1

Sub G1 cont

```
Lys Arg Gln Ala His Glu Ala Trp Lys Ala Thr His Val Gly His Leu
            105                 110                 115

GGC GCG GGC GTG CAC TGG GCG GAG GAC CGC CTG GCC GAC GCG TTC GAC        1038
Gly Ala Gly Val His Trp Ala Glu Asp Arg Leu Ala Asp Ala Phe Asp
        120                 125                 130

GTG CCC CAC CGC GAG GAG CGC GCC CGG GCC AAC GGC CTG ACG GAG CTG        1086
Val Pro His Arg Glu Glu Arg Ala Arg Ala Asn Gly Leu Thr Glu Leu
    135                 140                 145

GAC TCC GCG GAG GCG CTG GCC AAG GCG CTG GGG CTG AGC GTC TCC AAG        1134
Asp Ser Ala Glu Ala Leu Ala Lys Ala Leu Gly Leu Ser Val Ser Lys
150                 155                 160                 165

CTC CGC TGG TTC GCG TTC CAC CGG GAG GTC GAC ACG GCC ACG CAC TAC        1182
Leu Arg Trp Phe Ala Phe His Arg Glu Val Asp Thr Ala Thr His Tyr
                170                 175                 180

GTG AGC TGG ACC ATT CCG AAG CGG GAC GGC AGC AAG CGC ACG ATT ACG        1230
Val Ser Trp Thr Ile Pro Lys Arg Asp Gly Ser Lys Arg Thr Ile Thr
            185                 190                 195

TCC CCC AAG CCT GAG CTG AAG GCA GCG CAG CGC TGG GTG CTG TCC AAC        1278
Ser Pro Lys Pro Glu Leu Lys Ala Ala Gln Arg Trp Val Leu Ser Asn
        200                 205                 210

GTC GTG GAG CGG CTG CCG GTC CAC GGC GCC GCC CAC GGC TTC GTG GCG        1326
Val Val Glu Arg Leu Pro Val His Gly Ala Ala His Gly Phe Val Ala
    215                 220                 225

GGA CGC TCC ATC CTC ACC AAC GCG CTG GCC CAC CAG GGC GCG GAC GTC        1374
Gly Arg Ser Ile Leu Thr Asn Ala Leu Ala His Gln Gly Ala Asp Val
230                 235                 240                 245

GTG GTC AAG GTG GAC CTC AAG GAC TTC TTC CCC TCC GTC ACC TGG CGC        1422
Val Val Lys Val Asp Leu Lys Asp Phe Phe Pro Ser Val Thr Trp Arg
                250                 255                 260

CGG GTG AAG GGC CTG TTG CGC AAG GGC GGC CTG CGG GAG GGC ACG TCC        1470
Arg Val Lys Gly Leu Leu Arg Lys Gly Gly Leu Arg Glu Gly Thr Ser
            265                 270                 275

ACG CTG CTG TCC CTC CTC TCC ACG GAA GCG CCG CGG GAG GCG GTC CAG        1518
Thr Leu Leu Ser Leu Leu Ser Thr Glu Ala Pro Arg Glu Ala Val Gln
        280                 285                 290

TTC CGC GGC AAG CTC CTG CAC GTC GCC AAG GGC CCG CGC GCC CTG CCC        1566
Phe Arg Gly Lys Leu Leu His Val Ala Lys Gly Pro Arg Ala Leu Pro
    295                 300                 305

CAG GGC GCC CCC ACG TCG CCC GGC ATC ACC AAC GCG CTC TGC CTG AAG        1614
Gln Gly Ala Pro Thr Ser Pro Gly Ile Thr Asn Ala Leu Cys Leu Lys
310                 315                 320                 325
```

```
CTC GAC AAG CGG CTG TCC GCC CTC GCG AAG CGG CTG GGC TTC ACC TAC        1662
Leu Asp Lys Arg Leu Ser Ala Leu Ala Lys Arg Leu Gly Phe Thr Tyr
                330                 335                 340

ACG CGC TAC GCG GAC GAC CTG ACC TTC TCC TGG ACG AAG GCG AAG CAG        1710
Thr Arg Tyr Ala Asp Asp Leu Thr Phe Ser Trp Thr Lys Ala Lys Gln
            345                 350                 355

CCC AAG CCG CGG CGG ACG CAG CGT CCC CCC GTC GCG GTC CTC CTG TCT        1758
Pro Lys Pro Arg Arg Thr Gln Arg Pro Pro Val Ala Val Leu Leu Ser
        360                 365                 370

CGC GTC CAG GAA GTG GTG GAG GCG GAG GGC TTC CGC GTG CAC CCG GAC        1806
Arg Val Gln Glu Val Val Glu Ala Glu Gly Phe Arg Val His Pro Asp
    375                 380                 385

AAG ACG CGC GTC GCC CGC AAG GGC ACG CGG CAG CGG GTC ACC GGG CTC        1854
Lys Thr Arg Val Ala Arg Lys Gly Thr Arg Gln Arg Val Thr Gly Leu
390                 395                 400                 405

GTC GTG AAT GCG GCG GGC AAG GAC GCG CCC GCG GCC CGA GTC CCG CGC        1902
Val Val Asn Ala Ala Gly Lys Asp Ala Pro Ala Ala Arg Val Pro Arg
                410                 415                 420

GAC GTC GTC CGC CAG CTC CGC GCC GCC ATC CAC AAC CGG AAG AAG GGC        1950
Asp Val Val Arg Gln Leu Arg Ala Ala Ile His Asn Arg Lys Lys Gly
            425                 430                 435

AAG CCG GGC CGC GAG GGC GAG TCG CTC GAG CAG CTC AAG GGC ATG GCC        1998
Lys Pro Gly Arg Glu Gly Glu Ser Leu Glu Gln Leu Lys Gly Met Ala
        440                 445                 450

GCC TTC ATC CAC ATG ACG GAC CCG GCC AAG GGC CGC GCC TTC CTG GCT        2046
Ala Phe Ile His Met Thr Asp Pro Ala Lys Gly Arg Ala Phe Leu Ala
    455                 460                 465

CAG CTC ACG GAG CTC GAG TCC ACG GCG AGC GCC GCT CCG CAG GCG GAG        2094
Gln Leu Thr Glu Leu Glu Ser Thr Ala Ser Ala Ala Pro Gln Ala Glu
470                 475                 480                 485

TGACGCTCAG CGCGCGTCCG TCGCCGACGT GCCGCGCGCC AGCAACGCCG CATTCAGCAA      2154

CTCCGTCAGC CGGCGCGGGT AC                                               2176
```

C4

Sub B' cont

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 263 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

```
Val Lys Leu Lys Pro Gly Met Asp Gly Pro Lys Val Lys Gln Trp Pro
1               5                   10                  15

Leu Thr Glu Glu Lys Ile Lys Ala Leu Val Glu Ile Cys Thr Glu Met
            20                  25                  30

Glu Lys Glu Gly Lys Ile Ser Lys Ile Gly Pro Glu Asn Pro Tyr Asn
        35                  40                  45

Thr Pro Val Phe Ala Ile Lys Lys Lys Asp Ser Thr Lys Trp Arg Lys
    50                  55                  60

Leu Val Asp Phe Arg Glu Leu Asn Lys Arg Thr Gln Asp Phe Trp Glu
65                  70                  75                  80

Val Gln Leu Gly Ile Pro His Pro Ala Gly Leu Lys Lys Lys Lys Ser
                85                  90                  95

Val Thr Val Leu Asp Val Gly Asp Ala Tyr Phe Ser Val Pro Leu Asp
            100                 105                 110

Glu Asp Phe Arg Lys Tyr Thr Ala Phe Thr Ile Pro Ser Ile Asn Asn
        115                 120                 125

Glu Thr Pro Gly Ile Arg Tyr Gln Tyr Asn Val Leu Pro Gln Gly Trp
    130                 135                 140

Lys Gly Ser Pro Ala Ile Phe Gln Ser Ser Met Thr Lys Ile Leu Glu
145                 150                 155                 160

Pro Phe Lys Lys Gln Asn Pro Asp Ile Val Ile Tyr Gln Tyr Met Asp
                165                 170                 175

Asp Leu Tyr Val Gly Ser Asp Leu Glu Ile Gly Gln His Arg Thr Lys
            180                 185                 190

Ile Glu Glu Leu Arg Gln His Leu Leu Arg Trp Gly Leu Thr Thr Pro
        195                 200                 205

Asp Lys Lys His Gln Lys Glu Pro Pro Phe Leu Trp Met Gly Tyr Glu
    210                 215                 220

Leu His Pro Asp Lys Trp Thr Val Gln Pro Ile Val Leu Pro Glu Lys
225                 230                 235                 240

Asp Ser Trp Thr Val Asn Asp Ile Gln Lys Leu Val Gly Lys Leu Asn
                245                 250                 255

Trp Ala Ser Gln Ile Tyr Pro
            260
```

Sub C1 cont.

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 263 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

```
Arg Pro Trp Ala Arg Thr Pro Pro Lys Ala Pro Arg Asn Gln Pro Val
1               5                   10                  15

Pro Phe Lys Pro Glu Arg Leu Gln Ala Leu Gln His Leu Val Arg Lys
            20                  25                  30

Ala Leu Glu Ala Gly His Ile Glu Pro Tyr Thr Gly Pro Gly Asn Asn
        35                  40                  45

Pro Val Phe Pro Val Lys Lys Ala Asn Gly Thr Trp Arg Phe Ile His
    50                  55                  60

Asp Leu Arg Ala Thr Asn Ser Leu Thr Ile Asp Leu Ser Ser Ser Ser
65                  70                  75                  80

Pro Gly Pro Pro Asp Leu Ser Ser Leu Pro Thr Thr Leu Ala His Leu
                85                  90                  95

Gln Thr Ile Asp Leu Arg Asp Ala Phe Phe Gln Ile Pro Leu Pro Lys
            100                 105                 110

Gln Phe Gln Pro Tyr Phe Ala Phe Thr Val Pro Gln Gln Cys Asn Tyr
        115                 120                 125

Gly Pro Gly Thr Arg Tyr Ala Trp Lys Val Leu Pro Gln Gly Phe Lys
    130                 135                 140

Asn Ser Pro Thr Leu Phe Glu Met Gln Leu Ala His Ile Leu Gln Pro
145                 150                 155                 160

Ile Arg Gln Ala Phe Pro Gln Cys Thr Ile Leu Gln Tyr Met Asp Asp
                165                 170                 175

Ile Leu Leu Ala Ser Pro Ser His Glu Asp Leu Leu Leu Leu Ser Glu
            180                 185                 190

Ala Thr Met Ala Ser Leu Ile Ser His Gly Leu Pro Val Ser Glu Asn
        195                 200                 205

Lys Thr Gln Gln Thr Pro Gly Thr Ile Lys Phe Leu Gly Gln Ile Ile
    210                 215                 220
```

C1

Sub Ent

```
Ser Pro Asn His Leu Thr Tyr Asp Ala Val Pro Thr Val Pro Ile Arg
225                 230                 235                 240

Ser Arg Trp Ala Leu Pro Glu Leu Gln Ala Leu Leu Gly Glu Ile Gln
            245                 250                 255

Trp Val Ser Lys Gly Thr Pro
            260
```

(2) INFORMATION FOR SEQ ID NO:4:

```
    (i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 259 amino acids
        (B) TYPE: amino acid
        (D) TOPOLOGY: linear

   (ii) MOLECULE TYPE: protein


   (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

Asn Val Leu Tyr Arg Ile Gly Ser Asp Asn Gln Tyr Thr Gln Phe Thr
1               5                   10                  15

Ile Pro Lys Lys Gly Lys Gly Val Arg Thr Ile Ser Ala Pro Thr Asp
            20                  25                  30

Arg Leu Lys Asp Ile Gln Arg Arg Ile Cys Asp Leu Leu Ser Asp Cys
        35                  40                  45

Arg Asp Glu Ile Phe Ala Ile Arg Lys Ile Ser Asn Asn Tyr Ser Phe
    50                  55                  60

Gly Phe Glu Arg Gly Lys Ser Ile Ile Leu Asn Ala Tyr Lys His Arg
65                  70                  75                  80

Gly Lys Gln Ile Ile Leu Asn Ile Asp Leu Lys Asp Phe Phe Glu Ser
                85                  90                  95

Phe Asn Phe Gly Arg Val Arg Gly Tyr Phe Leu Ser Asn Gln Asp Phe
            100                 105                 110

Leu Leu Asn Pro Val Val Ala Thr Thr Leu Ala Lys Ala Ala Cys Tyr
        115                 120                 125

Asn Gly Thr Leu Pro Gln Gly Ser Pro Cys Ser Pro Ile Ile Ser Asn
    130                 135                 140

Leu Ile Cys Asn Ile Met Asp Met Arg Leu Ala Lys Leu Ala Lys Lys
145                 150                 155                 160

Tyr Gly Cys Thr Tyr Ser Arg Tyr Ala Asp Asp Ile Thr Ile Ser Thr
                165                 170                 175
```

Sub G1 cont

Asn Lys Asn Thr Phe Pro Leu Glu Met Ala Thr Val Gln Pro Glu Gly
180 185 190

Val Val Leu Gly Lys Val Leu Val Lys Glu Ile Glu Asn Ser Gly Phe
195 200 205

Glu Ile Asn Asp Ser Lys Thr Arg Leu Thr Tyr Lys Thr Ser Arg Gln
210 215 220

Glu Val Thr Gly Leu Thr Val Asn Arg Ile Val Asn Ile Asp Arg Cys
225 230 235 240

Tyr Tyr Lys Lys Thr Arg Ala Leu Ala His Ala Leu Tyr Arg Thr Gly
245 250 255

Glu Tyr Lys

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 266 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

Ala Phe His Arg Glu Val Asp Thr Ala Thr His Tyr Val Ser Trp Thr
1 5 10 15

Ile Pro Lys Arg Asp Gly Ser Lys Arg Thr Ile Thr Ser Pro Lys Pro
20 25 30

Glu Leu Lys Ala Ala Gln Arg Trp Val Leu Ser Asn Val Val Glu Arg
35 40 45

Leu Pro Val His Gly Ala Ala His Gly Phe Val Ala Gly Arg Ser Ile
50 55 60

Leu Thr Asn Ala Leu Ala His Gln Gly Ala Asp Val Val Val Lys Val
65 70 75 80

Asp Leu Lys Asp Phe Phe Pro Ser Val Thr Trp Arg Arg Val Lys Gly
85 90 95

Leu Leu Arg Lys Gly Gly Leu Arg Glu Gly Thr Ser Thr Leu Leu Ser
100 105 110

Leu Leu Ser Thr Glu Ala Pro Arg Glu Ala Val Gln Phe Arg Gly Lys
115 120 125

C4

Sub G cont'

```
Leu Leu His Val Ala Lys Gly Pro Arg Ala Leu Pro Gln Gly Ala Pro
    130                 135                 140

Thr Ser Pro Gly Ile Thr Asn Ala Leu Cys Leu Lys Leu Asp Lys Arg
145                 150                 155                 160

Leu Ser Ala Leu Ala Lys Arg Leu Gly Phe Thr Tyr Thr Arg Tyr Ala
                165                 170                 175

Asp Asp Leu Thr Phe Ser Trp Thr Lys Ala Lys Gln Pro Lys Pro Arg
            180                 185                 190

Arg Thr Gln Arg Pro Pro Val Ala Val Leu Leu Ser Arg Val Gln Glu
        195                 200                 205

Val Val Glu Ala Glu Gly Phe Arg Val His Pro Asp Lys Thr Arg Val
    210                 215                 220

Ala Arg Lys Gly Thr Arg Gln Arg Val Thr Gly Leu Val Val Asn Ala
225                 230                 235                 240

Ala Gly Lys Asp Ala Pro Ala Ala Arg Val Pro Arg Asp Val Val Arg
                245                 250                 255

Gln Leu Arg Ala Ala Ile His Asn Arg Lys
            260                 265
```

Sub G cont

09508031.030601

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 111 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

```
Pro Thr Pro Glu Leu Thr Ala Pro Ser Ser Asp Ala Ala Ala Lys Arg
1               5                   10                  15

Glu Ala Arg Arg Leu Ala His Glu Ala Leu Leu Val Arg Ala Lys Ala
            20                  25                  30

Ile Asp Glu Ala Gly Gly Ala Asp Asp Trp Val Gln Ala Gln Leu Val
        35                  40                  45

Ser Lys Gly Leu Ala Val Glu Asp Leu Asp Phe Ser Ser Ala Ser Glu
    50                  55                  60

Lys Asp Lys Lys Ala Trp Lys Glu Lys Lys Lys Ala Glu Ala Thr Glu
65                  70                  75                  80
```

```
Arg Arg Ala Leu Lys Arg Gln Ala His Glu Ala Trp Lys Ala Thr His
                85                  90                  95

Val Gly His Leu Gly Ala Gly Val His Trp Ala Glu Asp Arg Leu
            100                 105                 110
```

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 110 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

```
Pro Asp Pro Asp Met Thr Arg Val Thr Asn Ser Pro Ser Leu Gln Ala
1               5                   10                  15

His Leu Gln Ala Leu Tyr Leu Val Gln His Glu Val Trp Arg Pro Leu
            20                  25                  30

Ala Ala Ala Tyr Gln Glu Gln Leu Asp Arg Pro Val Val Pro His Pro
        35                  40                  45

Tyr Arg Val Gly Asp Thr Val Trp Val Arg Arg His Gln Thr Lys Asn
    50                  55                  60

Leu Glu Pro Arg Trp Lys Gly Pro Tyr Thr Val Leu Leu Thr Thr Pro
65                  70                  75                  80

Thr Ala Leu Lys Val Asp Gly Ile Ala Ala Trp Ile His Ala Ala His
                85                  90                  95

Val Lys Ala Ala Asp Pro Gly Gly Gly Pro Ser Ser Arg Leu
            100                 105                 110
```

Sub G cont'

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 75 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

```
Gly Lys Asp Ala Pro Ala Ala Arg Val Pro Arg Asp Val Val Arg Gln
```

```
1               5                   10                  15

Leu Arg Ala Ala Ile His Asn Arg Lys Lys Gly Lys Pro Gly Arg Glu
            20                  25                  30

Gly Glu Ser Leu Glu Gln Leu Lys Gly Met Ala Ala Phe Ile His Met
        35                  40                  45

Thr Asp Pro Ala Lys Gly Arg Ala Phe Leu Ala Gln Leu Thr Glu Leu
    50                  55                  60

Glu Ser Thr Ala Ser Ala Ala Pro Gln Ala Glu
65                  70                  75
```

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 66 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

```
Gly Lys Glu Gly His Ser Ala Arg Gln Cys Arg Ala Pro Arg Arg Gln
1               5                   10                  15

Gly Cys Trp Lys Cys Gly Lys Pro Gly His Ile Met Thr Asn Cys Pro
            20                  25                  30

Asp Arg Gln Ala Gly Phe Leu Gly Leu Gly Pro Trp Gly Lys Lys Pro
        35                  40                  45

Arg Asn Phe Pro Val Ala Gln Val Pro Gln Gly Leu Thr Pro Thr Ala
    50                  55                  60

Pro Pro
65
```

Sub G' cont

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 68 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

```
Gly Pro Arg Ala Leu Pro Gln Gly Ala Pro Thr Ser Pro Gly Ile Thr
1               5                   10                  15

Asn Ala Leu Cys Leu Lys Leu Asp Lys Arg Leu Ser Ala Leu Ala Lys
            20                  25                  30

Arg Leu Gly Phe Thr Tyr Thr Arg Tyr Ala Asp Asp Leu Thr Phe Ser
        35                  40                  45

Trp Thr Lys Ala Lys Gln Pro Lys Pro Arg Arg Thr Gln Arg Pro Pro
    50                  55                  60

Val Ala Val Leu
65
```

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 68 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

C4

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

```
Tyr Asn Gly Thr Leu Pro Gln Gly Ser Pro Cys Ser Pro Ile Ile Ser
1               5                   10                  15

Asn Leu Ile Cys Asn Ile Met Asp Met Arg Leu Ala Lys Leu Ala Lys
            20                  25                  30

Lys Tyr Gly Cys Thr Tyr Ser Arg Tyr Ala Asp Asp Ile Thr Ile Ser
        35                  40                  45

Thr Asn Lys Asn Thr Phe Pro Leu Glu Met Ala Thr Val Gln Pro Glu
    50                  55                  60

Gly Val Val Leu
65
```

Sub G1 cont

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 68 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

Tyr Lys Asn Leu Leu Pro Gln Gly Ala Pro Ser Ser Pro Lys Leu Ala
1   5   10   15

Asn Leu Ile Cys Ser Lys Leu Asp Tyr Arg Ile Gln Gly Tyr Ala Gly
            20   25   30

Ser Arg Gly Leu Ile Tyr Thr Arg Tyr Ala Asp Asp Leu Thr Leu Ser
        35   40   45

Ala Gln Ser Met Lys Lys Val Val Lys Ala Arg Asp Phe Leu Phe Ser
    50   55   60

Ile Ile Pro Ser
65

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 67 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

Tyr Gln Tyr Asn Val Leu Pro Gln Gly Trp Lys Gly Ser Pro Ala Ile
1   5   10   15

Phe Gln Ser Ser Met Thr Lys Ile Leu Glu Pro Phe Lys Lys Gln Asn
            20   25   30

Pro Asp Ile Val Ile Tyr Gln Tyr Met Asp Asp Leu Tyr Val Gly Ser
        35   40   45

Asp Leu Glu Ile Gly Gln His Arg Thr Lys Ile Glu Glu Leu Arg Gln
    50   55   60

His Leu Leu
65

Sub G' cont

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 66 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

```
Tyr Ala Trp Lys Val Leu Pro Gln Gly Phe Lys Asn Ser Pro Thr Leu
1               5                   10                  15

Phe Glu Met Gln Leu Ala His Ile Leu Gln Pro Ile Arg Gln Ala Phe
            20                  25                  30

Pro Gln Cys Thr Ile Leu Gln Tyr Met Asp Asp Ile Leu Leu Ala Ser
        35                  40                  45

Pro Ser His Glu Asp Leu Leu Leu Leu Ser Glu Ala Thr Met Ala Ser
    50                  55                  60

Leu Ile
65
```

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 65 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

```
Leu Thr Trp Thr Arg Leu Pro Gln Gly Phe Lys Asn Ser Pro Thr Leu
1               5                   10                  15

Phe Asp Glu Ala Leu His Arg Asp Leu Ala Asp Phe Arg Ile Gln His
            20                  25                  30

Pro Asp Leu Ile Leu Leu Gln Tyr Val Asp Asp Leu Leu Leu Ala Ala
        35                  40                  45

Thr Ser Glu Leu Asp Cys Gln Gln Gly Thr Arg Ala Leu Leu Gln Thr
    50                  55                  60

Leu
65
```

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 65 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

```
Phe Gln Trp Lys Val Leu Pro Gln Gly Met Thr Cys Ser Pro Thr Ile
1               5                   10                  15

Cys Gln Leu Val Val Gly Gln Val Leu Glu Pro Leu Arg Leu Lys His
            20                  25                  30

Pro Ser Leu Cys Met Leu His Tyr Met Asp Asp Leu Leu Leu Ala Ala
        35                  40                  45

Ser Ser His Asp Gly Leu Glu Ala Ala Gly Glu Glu Val Ile Ser Thr
    50                  55                  60

Leu
65
```

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 65 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

```
Phe Ala Trp Arg Val Leu Pro Gln Gly Phe Ile Asn Ser Pro Ala Leu
1               5                   10                  15

Phe Glu Arg Ala Leu Gln Glu Pro Leu Arg Gln Val Ser Ala Ala Phe
            20                  25                  30

Ser Gln Ser Leu Leu Val Ser Tyr Met Asp Asp Ile Leu Tyr Ala Ser
        35                  40                  45

Pro Thr Glu Glu Gln Arg Ser Gln Cys Tyr Gln Ala Leu Ala Ala Arg
    50                  55                  60

Leu
65
```

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 61 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

```
Ile Ala Thr Asn Gly Val Pro Gln Gly Ala Ser Thr Ser Cys Gly Leu
1               5                   10                  15

Ala Thr Tyr Asn Val Leu Glu Leu Phe Leu Arg Tyr Asp Glu Leu Ile
            20                  25                  30

Met Tyr Ala Asp Asp Gly Ile Leu Cys Arg Gln Asp Pro Ser Thr Pro
        35                  40                  45

Asp Phe Ser Val. Glu Glu Ala Gly Val Val Gln Glu Pro
    50                  55                  60
```

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 61 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

```
Tyr Glu Tyr Leu Arg Met Pro Phe Gly Leu Lys Asn Ala Pro Ala Thr
1               5                   10                  15

Phe Gln Arg Cys Met Asn Asp Ile Leu Arg Pro Leu Leu Asn Lys His
            20                  25                  30

Cys Leu Val Tyr Leu Asp Asp Ile Ile Val Phe Ser Thr Ser Leu Asp
        35                  40                  45

Glu His Leu Gln Ser Leu Gly Leu Val Phe Glu Lys Leu
    50                  55                  60
```

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 61 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

```
Tyr Glu Phe Cys Arg Leu Pro Phe Gly Leu Arg Asn Ala Ser Ser Ile
1               5                   10                  15

Phe Gln Arg Ala Leu Asp Asp Val Leu Arg Glu Gln Ile Gly Lys Ile
            20                  25                  30

Cys Tyr Val Tyr Val Asp Asp Val Ile Ile Phe Ser Glu Asn Glu Ser
        35                  40                  45

Asp His Val Arg His Ile Asp Thr Val Leu Lys Cys Leu
    50                  55                  60
```

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 63 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

```
Cys Lys Leu Asn Lys Ala Ile Tyr Gly Leu Lys Gln Ala Ala Arg Cys
1               5                   10                  15

Trp Phe Arg Cys Ile Tyr Ile Leu Asp Lys Gly Asn Ile Asn Glu Asn
            20                  25                  30

Ile Tyr Val Leu Leu Tyr Val Asp Asp Val Val Ile Ala Thr Gly Asp
        35                  40                  45

Met Thr Arg Met Asn Asn Phe Lys Arg Tyr Leu Met Glu Lys Phe
    50                  55                  60
```

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 62 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

```
Cys Leu Leu Lys Lys Ser Leu Tyr Gly Leu Lys Gln Ser Pro Arg Gln
1               5                   10                  15
```

Trp Asn Ala Cys Val Tyr Val Lys Gln Val Ser Glu Gln Glu His Leu
            20   25   30

Tyr Leu Leu Leu Tyr Val Asp Asp Met Leu Ile Ala Gly Lys Ser Lys
        35   40   45

Ser Glu Ile Asn Lys Val Lys Glu Gln Leu Ser Met Glu Phe
    50   55   60

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 63 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

Ile Arg Leu Lys Lys Ser Leu Tyr Glu Leu Lys Gln Ser Gly Ala Asn
1   5   10   15

Trp Tyr Glu Glu Val Arg Gly Trp Ser Cys Val Phe Lys Asn Ser Gln
            20   25   30

Val Thr Ile Cys Leu Phe Val Asp Asp Met Val Leu Phe Ser Lys Asn
        35   40   45

Leu Asn Ser Asn Lys Arg Ile Ile Glu Lys Leu Lys Met Gln Tyr
    50   55   60

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 58 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(B) LOCATION: 15
(D) OTHER INFORMATION: /note= "The 2' position of this nucleotide is linked to the 5' position of nucleotide number 1 of SEQ ID NO: 25 of this application."

(ix) FEATURE:
(A) NAME/KEY: misc_binding
(B) LOCATION: 52..58

(D) OTHER INFORMATION: /note= "This region can hydrogen bond to nucleotides 61-67 of SEQ ID NO: 25 of this application."

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

```
CACGCAUGUA GGCAGAUUUG UUGGUUGUGA AUCGCAACCA GUGGCCUUAA UGGCAGGA          58
```

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 67 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ix) FEATURE:
(A) NAME/KEY: misc_feature
(B) LOCATION: 1
(D) OTHER INFORMATION: /note= "The 5' position of this nucleotide is linked to the 2' position of nucleotide number 15 of SEQ ID NO: 24 of this application."

(ix) FEATURE:
(A) NAME/KEY: misc_binding
(B) LOCATION: 61..67
(D) OTHER INFORMATION: /note= "This region can hydrogen bond to nucleotides 52-58 of SEQ ID NO: 24 of this application."

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

```
TCCTTCGCAC AGCACACCTG CCGTATAGCT CTGAATCAAG GATTTTAGGG AGGCGATTCC        60

TCCTGCC                                                                  67
```

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2423 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 418..2175

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

```
TGGCCATTNA GATACGGATT TTCACTTCCT TGACAGTGCA TGACTATGCT GCATGAAATN      60

GCATGATCGA TTGAGGATCG TCTTTGCTCA GATCCGCCAG AACTGGCGGG CTTTTGCTCA     120

TGTCATGCAT GTGCATGAAA ACCACTGCAT AAAGCGGGCA GGCGTGGCGG GGATACGAGC     180

GCGCGCTATC ACCGAAAATA GCCAAAATAC TTCTGGAAAA CAGAAAGTTG AAGTGATATG     240

TTCATAAACA CGCATGTAGG CAGATTTGTT GGTTGTGAAT CGCAACCAGT GGCCTTAATG     300

GCAGGAGGAA TCGCCTCCCT AAAATCCTTG ATTCAGAGCT ATACGGCAGG TGTGCTGTGC     360

GAAGGAGTGC CTGCATGCGT TTCTCCTTGG CCTTTTTTCC TCTGGGATGA AGAAGAA        417

ATG ACA AAA ACA TCT AAA CTT GAC GCA CTT AGG GCT GCT ACT TCA CGT       465
Met Thr Lys Thr Ser Lys Leu Asp Ala Leu Arg Ala Ala Thr Ser Arg
  1               5                  10                  15

GAA GAC TTG GCT AAA ATT TTA GAT ATT AAG TTG GTA TTT TTA ACT AAC       513
Glu Asp Leu Ala Lys Ile Leu Asp Ile Lys Leu Val Phe Leu Thr Asn
             20                  25                  30

GTT CTA TAT AGA ATC GGC TCG GAT AAT CAA TAC ACT CAA TTT ACA ATA       561
Val Leu Tyr Arg Ile Gly Ser Asp Asn Gln Tyr Thr Gln Phe Thr Ile
         35                  40                  45

CCG AAG AAA GGA AAA GGG GTA AGG ACT ATT TCT GCA CCT ACA GAC CGG       609
Pro Lys Lys Gly Lys Gly Val Arg Thr Ile Ser Ala Pro Thr Asp Arg
     50                  55                  60

TTG AAG GAC ATC CAA CGA AGA ATA TGT GAC TTA CTT TCT GAT TGT AGA       657
Leu Lys Asp Ile Gln Arg Arg Ile Cys Asp Leu Leu Ser Asp Cys Arg
 65                  70                  75                  80

GAT GAG ATC TTT GCT ATA AGG AAA ATT AGT AAC AAC TAT TCC TTT GGT       705
Asp Glu Ile Phe Ala Ile Arg Lys Ile Ser Asn Asn Tyr Ser Phe Gly
                 85                  90                  95

TTT GAG AGG GGA AAA TCA ATA ATC CTA AAT GCT TAT AAG CAT AGA GGC       753
Phe Glu Arg Gly Lys Ser Ile Ile Leu Asn Ala Tyr Lys His Arg Gly
            100                 105                 110

AAA CAA ATA ATA TTA AAT ATA GAT CTT AAG GAT TTT TTT GAA AGC TTT       801
Lys Gln Ile Ile Leu Asn Ile Asp Leu Lys Asp Phe Phe Glu Ser Phe
        115                 120                 125

AAT TTT GGA CGA GTT AGA GGA TAT TTT CTT TCC AAT CAG GAT TTT TTA       849
Asn Phe Gly Arg Val Arg Gly Tyr Phe Leu Ser Asn Gln Asp Phe Leu
    130                 135                 140

TTA AAT CCT GTG GTG GCA ACG ACA CTT GCA AAA GCT GCA TGC TAT AAT       897
Leu Asn Pro Val Val Ala Thr Thr Leu Ala Lys Ala Ala Cys Tyr Asn
```

```
145                 150                 155                 160

GGA ACC CTC CCC CAA GGA AGT CCA TGT TCT CCT ATT ATC TCA AAT CTA       945
Gly Thr Leu Pro Gln Gly Ser Pro Cys Ser Pro Ile Ile Ser Asn Leu
                165                 170                 175

ATT TGC AAT ATT ATG GAT ATG AGA TTA GCT AAG CTG GCT AAA AAA TAT       993
Ile Cys Asn Ile Met Asp Met Arg Leu Ala Lys Leu Ala Lys Lys Tyr
            180                 185                 190

GGA TGT ACT TAT AGC AGA TAT GCT GAT GAT ATA ACA ATT TCT ACA AAT      1041
Gly Cys Thr Tyr Ser Arg Tyr Ala Asp Asp Ile Thr Ile Ser Thr Asn
        195                 200                 205

AAA AAT ACA TTT CCG TTA GAA ATG GCT ACT GTG CAA CCT GAA GGG GTT      1089
Lys Asn Thr Phe Pro Leu Glu Met Ala Thr Val Gln Pro Glu Gly Val
    210                 215                 220

GTT TTG GGA AAA GTT TTG GTA AAA GAA ATA GAA AAC TCT GGA TTC GAA      1137
Val Leu Gly Lys Val Leu Val Lys Glu Ile Glu Asn Ser Gly Phe Glu
225                 230                 235                 240

ATA AAT GAT TCA AAG ACT AGG CTT ACG TAT AAG ACA TCA AGG CAA GAA      1185
Ile Asn Asp Ser Lys Thr Arg Leu Thr Tyr Lys Thr Ser Arg Gln Glu
                245                 250                 255

GTA ACG GGA CTT ACA GTT AAC AGA ATC GTT AAT ATT GAT AGA TGT TAT      1233
Val Thr Gly Leu Thr Val Asn Arg Ile Val Asn Ile Asp Arg Cys Tyr
            260                 265                 270

TAT AAA AAA ACT CGG GCG TTG GCA CAT GCT TTG TAT CGT ACA GGT GAA      1281
Tyr Lys Lys Thr Arg Ala Leu Ala His Ala Leu Tyr Arg Thr Gly Glu
        275                 280                 285

TAT AAA GTG CCA GAT GAA AAT GGT GTT TTA GTT TCA GGA GGT CTG GAT      1329
Tyr Lys Val Pro Asp Glu Asn Gly Val Leu Val Ser Gly Gly Leu Asp
    290                 295                 300

AAA CTT GAG GGG ATG TTT GGT TTT ATT GAT CAA GTT GAT AAG TTT AAC      1377
Lys Leu Glu Gly Met Phe Gly Phe Ile Asp Gln Val Asp Lys Phe Asn
305                 310                 315                 320

AAT ATA AAG AAA AAA CTG AAC AAG CAA CCT GAT AGA TAT GTA TTG ACT      1425
Asn Ile Lys Lys Lys Leu Asn Lys Gln Pro Asp Arg Tyr Val Leu Thr
                325                 330                 335

AAT GCG ACT TTG CAT GGT TTT AAA TTA AAG TTG AAT GCG CGA GAA AAA      1473
Asn Ala Thr Leu His Gly Phe Lys Leu Lys Leu Asn Ala Arg Glu Lys
            340                 345                 350

GCA TAT AGT AAA TTT ATT TAC TAT AAA TTT TTT CAT GGC AAC ACC TGT      1521
Ala Tyr Ser Lys Phe Ile Tyr Tyr Lys Phe Phe His Gly Asn Thr Cys
        355                 360                 365
```

```
CCT ACG ATA ATT ACA GAA GGG AAG ACT GAT CGG ATA TAT TTG AAG GCT      1569
Pro Thr Ile Ile Thr Glu Gly Lys Thr Asp Arg Ile Tyr Leu Lys Ala
    370                 375                 380

GCT TTG CAT TCT TTG GAG ACA TCA TAT CCT GAG TTG TTT AGA GAA AAA      1617
Ala Leu His Ser Leu Glu Thr Ser Tyr Pro Glu Leu Phe Arg Glu Lys
385                 390                 395                 400

ACA GAT AGT AAA AAG AAA GAA ATA AAT CTT AAT ATA TTT AAA TCT AAT      1665
Thr Asp Ser Lys Lys Lys Glu Ile Asn Leu Asn Ile Phe Lys Ser Asn
                405                 410                 415

GAA AAG ACC AAA TAT TTT TTA GAT CTT TCT GGG GGA ACT GCA GAT CTG      1713
Glu Lys Thr Lys Tyr Phe Leu Asp Leu Ser Gly Gly Thr Ala Asp Leu
            420                 425                 430

AAA AAA TTT GTA GAG CGT TAT AAA AAT AAT TAT GCT TCT TAT TAT GGT      1761
Lys Lys Phe Val Glu Arg Tyr Lys Asn Asn Tyr Ala Ser Tyr Tyr Gly
        435                 440                 445

TCT GTT CCA AAA CAG CCA GTG ATT ATG GTT CTT GAT AAT GAT ACA GGT      1809
Ser Val Pro Lys Gln Pro Val Ile Met Val Leu Asp Asn Asp Thr Gly
    450                 455                 460

CCA AGC GAT TTA CTT AAT TTT CTG CGC AAT AAA GTT AAA AGC TGC CCA      1857
Pro Ser Asp Leu Leu Asn Phe Leu Arg Asn Lys Val Lys Ser Cys Pro
465                 470                 475                 480

GAC GAT GTA ACT GAA ATG AGA AAG ATG AAA TAT ATT CAT GTT TTC TAT      1905
Asp Asp Val Thr Glu Met Arg Lys Met Lys Tyr Ile His Val Phe Tyr
                485                 490                 495

AAT TTA TAT ATA GTT CTC ACA CCA TTG AGT CCT TCC GGC GAA CAA ACT      1953
Asn Leu Tyr Ile Val Leu Thr Pro Leu Ser Pro Ser Gly Glu Gln Thr
            500                 505                 510

TCA ATG GAG GAT CTT TTC CCT AAA GAT ATT TTA GAT ATC AAG ATT GAT      2001
Ser Met Glu Asp Leu Phe Pro Lys Asp Ile Leu Asp Ile Lys Ile Asp
        515                 520                 525

GGT AAG AAA TTC AAC AAA AAT AAT GAT GGA GAC TCA AAA ACG GAA TAT      2049
Gly Lys Lys Phe Asn Lys Asn Asn Asp Gly Asp Ser Lys Thr Glu Tyr
    530                 535                 540

GGG AAG CAT ATT TTT TCC ATG AGG GTT GTT AGA GAT AAA AAG CGG AAA      2097
Gly Lys His Ile Phe Ser Met Arg Val Val Arg Asp Lys Lys Arg Lys
545                 550                 555                 560

ATA GAT TTT AAG GCA TTT TGT TGT ATT TTT GAT GCT ATA AAA GAT ATA      2145
Ile Asp Phe Lys Ala Phe Cys Cys Ile Phe Asp Ala Ile Lys Asp Ile
                565                 570                 575

AAG GAA CAT TAT AAA TTA ATG TTA AAT AGC TAATGAACAG CCCTAACGTT        2195
Lys Glu His Tyr Lys Leu Met Leu Asn Ser
```

```
              580                   585

ATGAACGCTA AGGCTGATTT TTCGTTAAAA TTTATATGGT TTGAATTGTA ATATATTATC    2255

TTCAAGCCAT TTATTTAATT CCTGCATCCT TTTCTGTAAG GGTATTAATT CGTTCCTCAC    2315

AAACACTAAA CTCGCTTTTT CCACATCCCC AAACCCCCCT AACATTATTC GGCATAATCC    2375

CCATCATTTG CGGTGGCACA CGATGCGCTG CCATCATGTC ATCGCGGC                 2423
```

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 546 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

```
Val Lys Leu Lys Pro Gly Met Asp Gly Pro Lys Val Lys Gln Trp Pro
1               5                   10                  15

Leu Thr Glu Glu Lys Ile Lys Ala Leu Val Glu Ile Cys Thr Glu Met
            20                  25                  30

Glu Lys Glu Gly Lys Ile Ser Lys Ile Gly Pro Glu Asn Pro Tyr Asn
        35                  40                  45

Thr Pro Val Phe Ala Ile Lys Lys Lys Asp Ser Thr Lys Trp Arg Lys
    50                  55                  60

Leu Val Asp Phe Arg Glu Leu Asn Lys Arg Thr Gln Asp Phe Trp Glu
65                  70                  75                  80

Val Gln Leu Gly Ile Pro His Pro Ala Gly Leu Lys Lys Lys Lys Ser
                85                  90                  95

Val Thr Val Leu Asp Val Gly Asp Ala Tyr Phe Ser Val Pro Leu Asp
            100                 105                 110

Glu Asp Phe Arg Lys Tyr Thr Ala Phe Thr Ile Pro Ser Ile Asn Asn
        115                 120                 125

Glu Thr Pro Gly Ile Arg Tyr Gln Tyr Asn Val Leu Pro Gln Gly Trp
    130                 135                 140

Lys Gly Ser Pro Ala Ile Phe Gln Ser Ser Met Thr Lys Ile Leu Glu
145                 150                 155                 160
```

```
Pro Phe Lys Lys Gln Asn Pro Asp Ile Val Ile Tyr Gln Tyr Met Asp
                165                 170                 175

Asp Leu Tyr Val Gly Ser Asp Leu Glu Ile Gly Gln His Arg Thr Lys
            180                 185                 190

Ile Glu Glu Leu Arg Gln His Leu Leu Arg Trp Gly Leu Thr Thr Pro
        195                 200                 205

Asp Lys Lys His Gln Lys Glu Pro Pro Phe Leu Trp Met Gly Tyr Glu
    210                 215                 220

Leu His Pro Asp Lys Trp Thr Val Gln Pro Ile Val Leu Pro Glu Lys
225                 230                 235                 240

Asp Ser Trp Thr Val Asn Asp Ile Gln Lys Leu Val Gly Lys Leu Asn
                245                 250                 255

Trp Ala Ser Gln Ile Tyr Pro Gly Ile Lys Val Arg Gln Leu Cys Lys
            260                 265                 270

Leu Leu Arg Gly Thr Lys Ala Leu Thr Glu Val Ile Pro Leu Thr Glu
        275                 280                 285

Glu Ala Glu Leu Glu Leu Ala Glu Asn Arg Glu Ile Leu Lys Glu Pro
    290                 295                 300

Val His Gly Val Tyr Tyr Asp Pro Ser Lys Asp Leu Ile Ala Glu Ile
305                 310                 315                 320

Gln Lys Gln Gly Gln Gly Gln Trp Thr Tyr Gln Ile Tyr Gln Glu Pro
                325                 330                 335

Phe Lys Asn Leu Lys Thr Gly Lys Tyr Ala Arg Met Arg Gly Ala His
            340                 345                 350

Thr Asn Asp Val Lys Gln Leu Thr Glu Ala Val Gln Lys Ile Thr Thr
        355                 360                 365

Glu Ser Ile Val Ile Trp Gly Lys Thr Pro Lys Phe Lys Leu Pro Ile
    370                 375                 380

Gln Lys Glu Thr Trp Glu Thr Trp Trp Thr Glu Tyr Trp Gln Ala Thr
385                 390                 395                 400

Trp Ile Pro Glu Trp Glu Phe Val Asn Thr Pro Pro Leu Val Lys Leu
                405                 410                 415

Trp Tyr Gln Leu Glu Lys Glu Pro Ile Val Gly Ala Glu Thr Phe Tyr
            420                 425                 430

Val Asp Gly Ala Ala Asn Arg Glu Thr Lys Leu Gly Lys Ala Gly Tyr
        435                 440                 445
```

```
Val Thr Asn Lys Gly Arg Gln Lys Val Val Pro Leu Thr Asn Thr Thr
    450                 455                 460

Asn Gln Lys Thr Glu Leu Gln Ala Ile Tyr Leu Ala Leu Gln Asp Ser
465                 470                 475                 480

Gly Leu Glu Val Asn Ile Val Thr Asp Ser Gln Tyr Ala Leu Gln Ile
                485                 490                 495

Ile Gln Ala Gln Pro Asp Lys Ser Glu Ser Glu Leu Val Asn Gln Ile
            500                 505                 510

Ile Glu Gln Leu Ile Lys Lys Glu Lys Val Tyr Leu Ala Trp Val Pro
        515                 520                 525

Ala His Lys Gly Ile Gly Gly Asn Glu Gln Val Asp Lys Leu Val Ser
    530                 535                 540

Ala Gly
545
```

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 578 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

```
Arg Pro Trp Ala Arg Thr Pro Pro Lys Ala Pro Arg Asn Gln Pro Val
1               5                   10                  15

Pro Phe Lys Pro Glu Arg Leu Gln Ala Leu Gln His Leu Val Arg Lys
            20                  25                  30

Ala Leu Glu Ala Gly His Ile Glu Pro Tyr Thr Gly Pro Gly Asn Asn
        35                  40                  45

Pro Val Phe Pro Val Lys Lys Ala Asn Gly Thr Trp Arg Phe Ile His
    50                  55                  60

Asp Leu Arg Ala Thr Asn Ser Leu Thr Ile Asp Leu Ser Ser Ser Ser
65                  70                  75                  80

Pro Gly Pro Pro Asp Leu Ser Ser Leu Pro Thr Thr Leu Ala His Leu
                85                  90                  95

Gln Thr Ile Asp Leu Arg Asp Ala Phe Phe Gln Ile Pro Leu Pro Lys
            100                 105                 110
```

```
Gln Phe Gln Pro Tyr Phe Ala Phe Thr Val Pro Gln Gln Cys Asn Tyr
        115                 120                 125

Gly Pro Gly Thr Arg Tyr Ala Trp Lys Val Leu Pro Gln Gly Phe Lys
    130                 135                 140

Asn Ser Pro Thr Leu Phe Glu Met Gln Leu Ala His Ile Leu Gln Pro
145                 150                 155                 160

Ile Arg Gln Ala Phe Pro Gln Cys Thr Ile Leu Gln Tyr Met Asp Asp
                165                 170                 175

Ile Leu Leu Ala Ser Pro Ser His Glu Asp Leu Leu Leu Leu Ser Glu
            180                 185                 190

Ala Thr Met Ala Ser Leu Ile Ser His Gly Leu Pro Val Ser Glu Asn
        195                 200                 205

Lys Thr Gln Gln Thr Pro Gly Thr Ile Lys Phe Leu Gly Gln Ile Ile
    210                 215                 220

Ser Pro Asn His Leu Thr Tyr Asp Ala Val Pro Thr Val Pro Ile Arg
225                 230                 235                 240

Ser Arg Trp Ala Leu Pro Glu Leu Gln Ala Leu Leu Gly Glu Ile Gln
                245                 250                 255

Trp Val Ser Lys Gly Thr Pro Thr Leu Arg Gln Pro Leu His Ser Leu
            260                 265                 270

Tyr Cys Ala Leu Gln Arg His Thr Asp Pro Arg Asp Gln Ile Tyr Leu
        275                 280                 285

Asn Pro Ser Gln Val Gln Ser Leu Val Gln Leu Arg Gln Ala Leu Ser
    290                 295                 300

Gln Asn Cys Arg Ser Arg Leu Val Gln Thr Leu Pro Leu Leu Gly Ala
305                 310                 315                 320

Ile Met Leu Thr Leu Thr Gly Thr Thr Thr Val Val Phe Gln Ser Lys
                325                 330                 335

Glu Gln Trp Pro Leu Val Trp Leu His Ala Pro Leu Pro His Thr Ser
            340                 345                 350

Gln Cys Pro Trp Gly Gln Leu Leu Ala Ser Ala Val Leu Leu Leu Asp
        355                 360                 365

Lys Tyr Thr Leu Gln Ser Tyr Gly Leu Leu Cys Gln Thr Ile His His
    370                 375                 380

Asn Ile Ser Thr Gln Thr Phe Asn Gln Phe Ile Gln Thr Ser Asp His
385                 390                 395                 400
```

```
Pro Ser Val Pro Ile Leu Leu His His Ser His Arg Phe Lys Asn Leu
                405                 410                 415

Gly Ala Gln Thr Gly Glu Leu Trp Asn Thr Phe Leu Lys Thr Ala Ala
            420                 425                 430

Pro Leu Ala Pro Val Lys Ala Leu Met Pro Val Phe Thr Leu Ser Pro
        435                 440                 445

Val Ile Ile Asn Thr Ala Pro Cys Leu Phe Ser Asp Gly Ser Thr Ser
    450                 455                 460

Arg Ala Ala Tyr Ile Leu Trp Asp Lys Gln Ile Leu Ser Gln Arg Ser
465                 470                 475                     480

Phe Pro Leu Pro Pro Pro His Lys Ser Ala Gln Arg Ala Glu Leu Leu
                485                 490                 495

Gly Leu Leu His Gly Leu Ser Ser Ala Arg Ser Trp Arg Cys Leu Asn
            500                 505                 510

Ile Phe Leu Asp Ser Lys Tyr Leu Tyr His Tyr Leu Arg Thr Leu Ala
        515                 520                 525

Leu Gly Thr Phe Gln Gly Arg Ser Ser Gln Ala Pro Phe Gln Ala Leu
    530                 535                 540

Leu Pro Arg Leu Leu Ser Arg Lys Val Val Tyr Leu His His Val Arg
545                 550                 555                     560

Ser His Thr Asn Leu Pro Asp Pro Ile Ser Arg Leu Asn Ala Leu Thr
                565                 570                 575

Asp Ala
```

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 555 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

```
Asn Val Leu Tyr Arg Ile Gly Ser Asp Asn Gln Tyr Thr Gln Phe Thr
1               5                   10                  15

Ile Pro Lys Lys Gly Lys Gly Val Arg Thr Ile Ser Ala Pro Thr Asp
            20                  25                  30
```

```
Arg Leu Lys Asp Ile Gln Arg Arg Ile Cys Asp Leu Leu Ser Asp Cys
        35                  40                  45

Arg Asp Glu Ile Phe Ala Ile Arg Lys Ile Ser Asn Asn Tyr Ser Phe
    50                  55                  60

Gly Phe Glu Arg Gly Lys Ser Ile Ile Leu Asn Ala Tyr Lys His Arg
65                  70                  75                  80

Gly Lys Gln Ile Ile Leu Asn Ile Asp Leu Lys Asp Phe Phe Glu Ser
                85                  90                  95

Phe Asn Phe Gly Arg Val Arg Gly Tyr Phe Leu Ser Asn Gln Asp Phe
            100                 105                 110

Leu Leu Asn Pro Val Val Ala Thr Thr Leu Ala Lys Ala Ala Cys Tyr
        115                 120                 125

Asn Gly Thr Leu Pro Gln Gly Ser Pro Cys Ser Pro Ile Ile Ser Asn
    130                 135                 140

Leu Ile Cys Asn Ile Met Asp Met Arg Leu Ala Lys Leu Ala Lys Lys
145                 150                 155                 160

Tyr Gly Cys Thr Tyr Ser Arg Tyr Ala Asp Asp Ile Thr Ile Ser Thr
                165                 170                 175

Asn Lys Asn Thr Phe Pro Leu Glu Met Ala Thr Val Gln Pro Glu Gly
            180                 185                 190

Val Val Leu Gly Lys Val Leu Val Lys Glu Ile Glu Asn Ser Gly Phe
        195                 200                 205

Glu Ile Asn Asp Ser Lys Thr Arg Leu Thr Tyr Lys Thr Ser Arg Gln
    210                 215                 220

Glu Val Thr Gly Leu Thr Val Asn Arg Ile Val Asn Ile Asp Arg Cys
225                 230                 235                 240

Tyr Tyr Lys Lys Thr Arg Ala Leu Ala His Ala Leu Tyr Arg Thr Gly
                245                 250                 255

Glu Tyr Lys Val Pro Asp Glu Asn Gly Val Leu Val Ser Gly Gly Leu
            260                 265                 270

Asp Lys Leu Glu Gly Met Phe Gly Phe Ile Asp Gln Val Asp Lys Phe
        275                 280                 285

Asn Asn Ile Lys Lys Lys Leu Asn Lys Gln Pro Asp Arg Tyr Val Leu
    290                 295                 300

Thr Asn Ala Thr Leu His Gly Phe Lys Leu Lys Leu Asn Ala Arg Glu
305                 310                 315                 320
```

```
Lys Ala Tyr Ser Lys Phe Ile Tyr Tyr Lys Phe Phe His Gly Asn Thr
                325                 330                 335

Cys Pro Thr Ile Ile Thr Glu Gly Lys Thr Asp Arg Ile Tyr Leu Lys
            340                 345                 350

Ala Ala Leu His Ser Leu Glu Thr Ser Tyr Pro Glu Leu Phe Arg Glu
        355                 360                 365

Lys Thr Asp Ser Lys Lys Lys Glu Ile Asn Leu Asn Ile Phe Lys Ser
    370                 375                 380

Asn Glu Lys Thr Lys Tyr Phe Leu Asp Leu Ser Gly Gly Thr Ala Asp
385                 390                 395                 400

Leu Lys Lys Phe Val Glu Arg Tyr Lys Asn Asn Tyr Ala Ser Tyr Tyr
                405                 410                 415

Gly Ser Val Pro Lys Gln Pro Val Ile Met Val Leu Asp Asn Asp Thr
            420                 425                 430

Gly Pro Ser Asp Leu Leu Asn Phe Leu Arg Asn Lys Val Lys Ser Cys
        435                 440                 445

Pro Asp Asp Val Thr Glu Met Arg Lys Met Lys Tyr Ile His Val Phe
    450                 455                 460

Tyr Asn Leu Tyr Ile Val Leu Thr Pro Leu Ser Pro Ser Gly Glu Gln
465                 470                 475                 480

Thr Ser Met Glu Asp Leu Phe Pro Lys Asp Ile Leu Asp Ile Lys Ile
                485                 490                 495

Asp Gly Lys Lys Phe Asn Lys Asn Asn Asp Gly Asp Ser Lys Thr Glu
            500                 505                 510

Tyr Gly Lys His Ile Phe Ser Met Arg Val Val Arg Asp Lys Lys Arg
        515                 520                 525

Lys Ile Asp Phe Lys Ala Phe Cys Cys Ile Phe Asp Ala Ile Lys Asp
    530                 535                 540

Ile Lys Glu His Tyr Lys Leu Met Leu Asn Ser
545                 550                 555
```

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
  (A) LENGTH: 243 amino acids
  (B) TYPE: amino acid
  (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

```
Arg Trp Phe Ser Phe His Arg Glu Val Asp Thr Gly Thr His Tyr Gln
1               5                   10                  15

Thr Trp Glu Ile Pro Lys Arg Asp Gly Gly Lys Arg Thr Leu Thr Ala
            20                  25                  30

Pro Lys Arg Glu Leu Lys Ala Val Gln Arg Trp Val Leu Ala Asn Val
        35                  40                  45

Val Glu Arg Leu Pro Val His Gly Ala Ala His Gly Phe Val Ala Gly
    50                  55                  60

Arg Ser Ile Leu Thr Asn Ala Leu Ala His Gln Gly Ala Asp Val Val
65                  70                  75                  80

Val Lys Val Asp Met Lys Asp Phe Phe Pro Ser Val Thr Trp Pro Arg
                85                  90                  95

Val Lys Gly Leu Leu Arg Lys Gly Gly Leu Pro Glu Asn Leu Ala Thr
            100                 105                 110

Leu Leu Ala Leu Leu Ser Thr Glu Ala Pro Arg Glu Val Val Arg Phe
        115                 120                 125

Arg Gly Glu Thr Leu Tyr Val Ala Lys Gly Pro Arg Ala Leu Pro Gln
    130                 135                 140

Gly Ala Pro Thr Ser Pro Ala Leu Thr Asn Ala Leu Cys Leu Arg Leu
145                 150                 155                 160

Asp Lys Arg Leu Ser Ala Leu Ser Lys Arg Leu Gly Phe Thr Tyr Thr
                165                 170                 175

Arg Tyr Ala Asp Asp Leu Thr Phe Ser Trp Arg Arg Ala Lys Lys Ser
            180                 185                 190

Arg Gln Lys Glu Leu Pro Leu Ala Asp Ala Pro Val Ala Leu Leu Leu
        195                 200                 205

Ala Arg Val Lys Gly Val Leu Glu Ala Glu Gly Phe Thr Leu His Pro
    210                 215                 220

Asp Lys Thr Arg Val Gln Arg Lys Gly Ser Arg Gln Arg Val Thr Gly
225                 230                 235                 240

Leu Val Val
```

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 241 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

```
Arg Trp Phe Ala Phe His Arg Glu Val Asp Thr Ala Thr His Tyr Val
1               5                   10                  15

Ser Trp Thr Ile Pro Lys Arg Asp Gly Ser Lys Arg Thr Ile Thr Ser
            20                  25                  30

Pro Lys Pro Glu Leu Lys Ala Ala Gln Arg Trp Val Leu Ser Asn Val
        35                  40                  45

Val Glu Arg Leu Pro Val His Gly Ala Ala His Gly Phe Val Ala Gly
    50                  55                  60

Arg Ser Ile Leu Thr Asn Ala Leu Ala His Gln Gly Ala Asp Val Val
65                  70                  75                  80

Val Lys Val Asp Leu Lys Asp Phe Phe Pro Ser Val Thr Trp Arg Arg
                85                  90                  95

Val Lys Gly Leu Leu Arg Lys Gly Gly Leu Arg Glu Gly Thr Ser Thr
            100                 105                 110

Leu Leu Ser Leu Leu Ser Thr Glu Ala Pro Arg Glu Ala Val Gln Phe
        115                 120                 125

Pro Arg Glu Leu Leu His Val Ala Lys Gly Pro Arg Ala Leu Pro Gln
    130                 135                 140

Gly Ala Pro Thr Ser Pro Gly Ile Thr Asn Ala Leu Cys Leu Lys Leu
145                 150                 155                 160

Asp Lys Arg Leu Ser Ala Leu Ala Lys Arg Leu Gly Phe Thr Tyr Thr
                165                 170                 175

Arg Tyr Ala Asp Asp Leu Thr Phe Ser Trp Thr Lys Ala Lys Gln Pro
            180                 185                 190

Lys Pro Arg Arg Thr Gln Arg Pro Pro Val Ala Val Leu Leu Ser Arg
        195                 200                 205

Val Gln Glu Val Val Glu Ala Glu Gly Phe Arg Val His Pro Asp Lys
    210                 215                 220

Thr Arg Val Ala Arg Lys Gly Thr Arg Gln Arg Val Thr Gly Leu Val
225                 230                 235                 240
```

Val

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 231 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

```
Arg His Tyr Ser Ile His Arg Pro Arg Glu Arg Val Arg His Tyr Val
1               5                   10                  15

Thr Phe Ala Val Pro Lys Arg Ser Gly Gly Val Arg Leu Leu His Ala
            20                  25                  30

Pro Lys Arg Arg Leu Lys Ala Leu Gln Arg Arg Met Leu Ala Leu Leu
        35                  40                  45

Val Ser Lys Leu Pro Val Ser Pro Gln Ala His Gly Phe Val Pro Gly
    50                  55                  60

Arg Ser Ile Lys Thr Gly Ala Ala Pro His Val Gly Arg Arg Val Val
65                  70                  75                  80

Leu Lys Leu Asp Leu Lys Asp Phe Phe Pro Ser Val Thr Phe Ala Arg
                85                  90                  95

Val Arg Gly Leu Leu Lys Ala Leu Gly Tyr Gly Tyr Pro Val Ala Ala
            100                 105                 110

Thr Leu Ala Val Leu Met Thr Glu Ser Glu Arg Gln Pro Val Glu Leu
        115                 120                 125

Glu Gly Ile Leu Phe His Val Pro Val Gly Pro Arg Val Cys Val Gln
    130                 135                 140

Gly Ala Pro Thr Ser Pro Ala Leu Cys Asn Ala Val Leu Leu Arg Leu
145                 150                 155                 160

Asp Arg Arg Leu Ala Gly Leu Ala Arg Arg Tyr Gly Tyr Thr Tyr Thr
                165                 170                 175

Arg Tyr Ala Asp Asp Leu Thr Phe Ser Gly Asp Asp Val Thr Ala Leu
            180                 185                 190

Glu Arg Val Arg Ala Leu Ala Ala Arg Tyr Val Gln Glu Glu Gly Phe
        195                 200                 205
```

```
Glu Val Asn Arg Glu Lys Thr Arg Val Gln Arg Arg Gly Gly Ala Gln
    210                 215                 220

Arg Val Thr Gly Val Thr Val
225                 230
```

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 234 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

```
Phe Leu Thr Asn Val Leu Tyr Arg Ile Gly Ser Asp Asn Gln Tyr Thr
1               5                   10                  15

Gln Phe Thr Ile Pro Lys Lys Gly Lys Gly Val Arg Thr Ile Ser Ala
            20                  25                  30

Pro Thr Asp Arg Leu Lys Asp Ile Gln Arg Arg Ile Cys Asp Leu Leu
        35                  40                  45

Ser Asp Cys Arg Asp Glu Ile Phe Ala Ile Arg Lys Ile Ser Asn Asn
    50                  55                  60

Tyr Ser Phe Gly Phe Glu Arg Gly Lys Ser Ile Ile Leu Asn Ala Tyr
65                  70                  75                  80

Lys His Arg Gly Lys Gln Ile Ile Leu Asn Ile Asp Leu Lys Asp Phe
                85                  90                  95

Phe Glu Ser Phe Asn Phe Gly Arg Val Arg Gly Tyr Phe Leu Ser Asn
            100                 105                 110

Gln Asp Phe Leu Leu Asn Pro Val Val Ala Thr Thr Leu Ala Lys Ala
        115                 120                 125

Ala Cys Tyr Asn Gly Thr Leu Pro Gln Gly Ser Pro Cys Ser Pro Ile
    130                 135                 140

Ile Ser Asn Leu Ile Cys Asn Ile Met Asp Met Arg Leu Ala Lys Leu
145                 150                 155                 160

Ala Lys Lys Tyr Gly Cys Thr Tyr Ser Arg Tyr Ala Asp Asp Ile Thr
                165                 170                 175

Ile Ser Thr Asn Lys Asn Thr Phe Pro Leu Glu Met Ala Thr Val Gln
            180                 185                 190
```

```
Pro Glu Gly Val Val Leu Gly Lys Val Leu Val Lys Glu Ile Glu Asn
        195                 200                 205

Ser Gly Phe Glu Ile Asn Asp Ser Lys Thr Arg Leu Thr Tyr Lys Thr
    210                 215                 220

Ser Arg Gln Glu Val Thr Gly Leu Thr Val
225                 230
```

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 215 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

```
Val Glu Thr Leu Arg Leu Leu Ile Tyr Thr Ala Asp Phe Arg Tyr Arg
1               5                   10                  15

Ile Tyr Thr Val Glu Lys Lys Gly Pro Glu Lys Arg Met Arg Thr Ile
            20                  25                  30

Tyr Gln Pro Ser Arg Glu Leu Lys Ala Leu Gln Gly Trp Val Leu Arg
        35                  40                  45

Asn Ile Leu Asp Lys Leu Ser Ser Ser Pro Phe Ser Ile Gly Phe Glu
    50                  55                  60

Lys His Gln Ser Ile Leu Asn Asn Ala Thr Pro His Ile Gly Ala Asn
65                  70                  75                  80

Phe Ile Leu Asn Ile Asp Leu Glu Asp Phe Phe Pro Ser Leu Thr Ala
                85                  90                  95

Asn Lys Val Phe Gly Val Phe His Ser Leu Gly Tyr Asn Arg Leu Ile
            100                 105                 110

Ser Ser Val Leu Thr Lys Ile Cys Cys Tyr Lys Asn Leu Leu Pro Gln
        115                 120                 125

Gly Ala Pro Ser Ser Pro Lys Leu Ala Asn Leu Ile Cys Ser Lys Leu
    130                 135                 140

Asp Tyr Arg Ile Gln Gly Tyr Ala Gly Ser Arg Gly Leu Ile Tyr Thr
145                 150                 155                 160

Arg Tyr Ala Asp Asp Leu Thr Leu Ser Ala Gln Ser Met Lys Lys Val
                165                 170                 175
```

```
Val Lys Ala Arg Asp Phe Leu Phe Ser Ile Ile Pro Ser Glu Gly Leu
            180                 185                 190

Val Ile Asn Ser Lys Lys Thr Cys Ile Ser Gly Pro Arg Ser Gln Arg
        195                 200                 205

Lys Val Thr Gly Leu Val Ile
    210                 215
```

(2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 230 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

```
Thr Lys Gly Phe Ala Ser Glu Val Met Arg Ser Pro Glu Pro Pro Lys
1               5                   10                  15

Lys Trp Asp Ile Ala Lys Lys Lys Gly Gly Met Arg Thr Ile Tyr His
            20                  25                  30

Pro Ser Ser Lys Val Lys Leu Ile Gln Tyr Trp Leu Met Asn Asn Val
        35                  40                  45

Phe Ser Lys Leu Pro Met His Asn Ala Ala Tyr Ala Phe Val Lys Asn
    50                  55                  60

Arg Ser Ile Lys Ser Asn Ala Leu Leu His Ala Glu Ser Lys Asn Lys
65                  70                  75                  80

Tyr Tyr Val Lys Ile Asp Leu Lys Asp Phe Phe Pro Ser Ile Lys Phe
                85                  90                  95

Thr Asp Phe Glu Tyr Ala Phe Thr Arg Tyr Arg Asp Arg Ile Glu Phe
            100                 105                 110

Thr Thr Glu Tyr Asp Leu Glu Leu Leu Gln Leu Ile Lys Thr Ile Cys
        115                 120                 125

Phe Ile Ser Asp Ser Thr Leu Pro Ile Gly Phe Pro Thr Ser Pro Leu
    130                 135                 140

Ile Ala Asn Phe Val Ala Arg Glu Leu Asp Glu Lys Leu Thr Gln Lys
145                 150                 155                 160

Leu Asn Ala Ile Asp Lys Leu Asn Ala Thr Tyr Thr Arg Tyr Ala Asp
                165                 170                 175
```

```
Asp Ile Ile Val Ser Thr Asn Met Lys Gly Ala Ser Lys Leu Ile Leu
            180                 185                 190

Asp Cys Phe Lys Arg Thr Met Lys Glu Ile Gly Pro Asp Phe Lys Ile
        195                 200                 205

Asn Ile Lys Lys Phe Lys Ile Cys Ser Ala Ser Gly Gly Ser Ile Val
    210                 215                 220

Val Thr Gly Leu Lys Val
225                 230
```

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 211 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

```
Ile Gln Arg Leu His Ala Leu Ser Asn His Ala Gly Arg His Tyr Arg
1               5                   10                  15

Arg Ile Ile Leu Ser Lys Arg His Gly Gly Gln Arg Leu Val Leu Ala
            20                  25                  30

Pro Asp Tyr Leu Leu Lys Thr Val Gln Arg Asn Ile Leu Lys Asn Val
        35                  40                  45

Leu Ser Gln Phe Pro Leu Ser Pro Phe Ala Thr Ala Tyr Arg Pro Gly
    50                  55                  60

Cys Pro Ile Val Ser Asn Ala Gln Pro His Cys Gln Gln Pro Gln Ile
65                  70                  75                  80

Leu Lys Leu Asp Ile Glu Asn Phe Phe Asp Ser Ile Ser Trp Leu Gln
                85                  90                  95

Val Trp Arg Val Phe Arg Gln Ala Gln Leu Pro Arg Asn Val Val Thr
            100                 105                 110

Met Leu Thr Trp Ile Cys Cys Tyr Asn Asp Ala Leu Pro Gln Gly Ala
        115                 120                 125

Pro Thr Ser Pro Ala Ile Ser Asn Leu Val Met Arg Arg Phe Asp Glu
    130                 135                 140

Arg Ile Gly Glu Trp Cys Gln Ala Arg Gly Ile Thr Tyr Thr Arg Tyr
145                 150                 155                 160
```

```
     55                  60                  65

CGC CGC TAC ACC CCG GGC CGG AAG AAG TGG ATG GAG GCC GCC GAG GCC       533
Arg Arg Tyr Thr Pro Gly Arg Lys Lys Trp Met Glu Ala Ala Glu Ala
 70                  75                  80                  85

CGG CGG CTG TTC TCC GCC ACG CTG CGC ACG CGG AAC CGG AAC CTG AGG       581
Arg Arg Leu Phe Ser Ala Thr Leu Arg Thr Arg Asn Arg Asn Leu Arg
                 90                  95                 100

GAC TTG CTG CCC GAC GAG GCA CAG CTG GCG CGC TAC GGC CTG CCG GTC       629
Asp Leu Leu Pro Asp Glu Ala Gln Leu Ala Arg Tyr Gly Leu Pro Val
            105                 110                 115

TGG CGC ACG GAA GAG GAC GTG GCA GCG GCC CTG GGC GTC TCG GTG GGC       677
Trp Arg Thr Glu Glu Asp Val Ala Ala Ala Leu Gly Val Ser Val Gly
        120                 125                 130

GTG CTC CGC CAC TAC AGC ATC CAC CGC CCG CGC GAG CGG GTG CGG CAC       725
Val Leu Arg His Tyr Ser Ile His Arg Pro Arg Glu Arg Val Arg His
    135                 140                 145

TAC GTG ACC TTC GCC GTG CCC AAG CGC TCC GGA GGC GTC CGG CTG CTG       773
Tyr Val Thr Phe Ala Val Pro Lys Arg Ser Gly Gly Val Arg Leu Leu
150                 155                 160                 165

CAT GCG CCC AAG CGG CGC CTG AAG GCC CTG CAA CGC CGG ATG CTG GCG       821
His Ala Pro Lys Arg Arg Leu Lys Ala Leu Gln Arg Arg Met Leu Ala
                170                 175                 180

CTC CTG GTG TCG AAG CTC CCC GTG AGT CCA CAG GCC CAT GGC TTC GTG       869
Leu Leu Val Ser Lys Leu Pro Val Ser Pro Gln Ala His Gly Phe Val
            185                 190                 195

CCC GGC CGC TCC ATC AAG ACG GGC GCC GCG CCG CAC GTG GGC CGG CGG       917
Pro Gly Arg Ser Ile Lys Thr Gly Ala Ala Pro His Val Gly Arg Arg
        200                 205                 210

GTG GTC CTG AAG CTG GAC CTG AAG GAC TTC TTC CCC TCC GTC ACC TTC       965
Val Val Leu Lys Leu Asp Leu Lys Asp Phe Phe Pro Ser Val Thr Phe
    215                 220                 225

GCG CGG GTG CGA GGG CTG CTC ATC GCC CTG GGC TAC GGC TAT CCC GTG      1013
Ala Arg Val Arg Gly Leu Leu Ile Ala Leu Gly Tyr Gly Tyr Pro Val
230                 235                 240                 245

GCG GCC ACG CTC GCG GTG CTG ATG ACG GAG TCC GAG CGC CAG CCC GTG      1061
Ala Ala Thr Leu Ala Val Leu Met Thr Glu Ser Glu Arg Gln Pro Val
                250                 255                 260

GAG CTG GAG GGC ATC CTC TTC CAC GTT CCC GTG GGC CCA CGC GTC TGC      1109
Glu Leu Glu Gly Ile Leu Phe His Val Pro Val Gly Pro Arg Val Cys
            265                 270                 275
```

```
Cys Asp Asp Met Thr Phe Ser Gly His Phe Asn Ala Arg Gln Val Lys
                165                 170                 175

Asn Lys Val Cys Gly Leu Leu Ala Glu Leu Gly Leu Ser Leu Asn Lys
            180                 185                 190

Arg Lys Gly Cys Leu Ile Ala Ala Cys Lys Arg Gln Gln Val Thr Gly
        195                 200                 205

Ile Val Val
    210
```

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1640 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 279..1559

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

```
CTCCGAGCCC GCCTCCGAGG ACGCGCTCGC GGCCCGGGCG GCGGGGGCGG ACGCGCGGCG      60

GCGGCCCACG GAGACGCTTG ACCCGGGAGA CGACGAATGA CGATAACGGC AGGTGCTCTC     120

GGGAGAGGCC AGGGCTCGCA GATGAGCCAT GAGTACCGCG GTGTTTCGCC GCGGGGGTGT     180

TCTGTCCCCA TCTCTTCGCC AGGGTCCCAG CGTACGCAAC GCAGGGAGCC CCGGGTCCAA     240

CGCCTCGCAG GTCGTCCCCT GGCCTCTTCC GGAGCACC ATG AGC TGG TTC GAC         293
                                          Met Ser Trp Phe Asp
                                            1               5

ACC ACC CTC TCC CGG CTC AAG GGG TTG TTC AGC CGT CCC GTG ACA CGA       341
Thr Thr Leu Ser Arg Leu Lys Gly Leu Phe Ser Arg Pro Val Thr Arg
                 10                  15                  20

AGC ACC ACC GGG CTG GAC GTG CCG CTG GAT GCC CAC GGA CGT CCC CAG       389
Ser Thr Thr Gly Leu Asp Val Pro Leu Asp Ala His Gly Arg Pro Gln
             25                  30                  35

GAC GTC GTG ACG GAG ACG GTC TCC ACG TCG GGC CCC CTG AAG CCA GGG       437
Asp Val Val Thr Glu Thr Val Ser Thr Ser Gly Pro Leu Lys Pro Gly
         40                  45                  50

CAC CTG CGA CAG GTC CGC CGG GAT GCG CGG CTG CTC CCC AAG GGC GTC       485
His Leu Arg Gln Val Arg Arg Asp Ala Arg Leu Leu Pro Lys Gly Val
```

```
GTG CAG GGC GCC CCC ACG AGC CCC GCC CTG TGC AAC GCG GTG CTG CTG       1157
Val Gln Gly Ala Pro Thr Ser Pro Ala Leu Cys Asn Ala Val Leu Leu
        280                 285                 290

CGA CTG GAC CGG CGG CTG GCG GGA CTG GCG CGT CGG TAC GGC TAC ACG       1205
Arg Leu Asp Arg Arg Leu Ala Gly Leu Ala Arg Arg Tyr Gly Tyr Thr
    295                 300                 305

TAC ACG CGC TAC GCG GAT GAC CTC ACC TTC TCC GGC GAC GAC GTC ACG       1253
Tyr Thr Arg Tyr Ala Asp Asp Leu Thr Phe Ser Gly Asp Asp Val Thr
310                 315                 320                 325

GCG CTG GAG CGA GTC CGC GCG CTG GCC GCG CGG TAC GTG CAG GAG GAA       1301
Ala Leu Glu Arg Val Arg Ala Leu Ala Ala Arg Tyr Val Gln Glu Glu
                330                 335                 340

GGC TTC GAG GTC AAC CGC GAG AAG ACC CGC GTG CAG CGC CGG GGC GGT       1349
Gly Phe Glu Val Asn Arg Glu Lys Thr Arg Val Gln Arg Arg Gly Gly
            345                 350                 355

GCC CAG CGC GTC ACT GGC GTC ACC GTG AAT ACG ACG CTG GGC TTG TCA       1397
Ala Gln Arg Val Thr Gly Val Thr Val Asn Thr Thr Leu Gly Leu Ser
        360                 365                 370

CGC GAG GAG CGG CCG CGG CTC CGG GCG ATG CTG CAC CAG GAG GCG CGG       1445
Arg Glu Glu Arg Pro Arg Leu Arg Ala Met Leu His Gln Glu Ala Arg
    375                 380                 385

TCG GAG GAC GTC GAG GCA CAC CGC GCG CAC CTC GAC GGC CTC CTG GCC       1493
Ser Glu Asp Val Glu Ala His Arg Ala His Leu Asp Gly Leu Leu Ala
390                 395                 400                 405

TAC GTG AAG ATG CTC AAC CCG GAG CAG GCG GAG CGG CTC GCT CGC CGG       1541
Tyr Val Lys Met Leu Asn Pro Glu Gln Ala Glu Arg Leu Ala Arg Arg
                410                 415                 420

CGC AAG CCG CGC GGG ACG TGAGCGAGGG CTCAGCTCCG GATGGGCCAG              1589
Arg Lys Pro Arg Gly Thr
            425

GGCCTGTCAC GCGTCCCGGC CTCCCAGTTG TCATGGCGGC CGTCCCAGTA C              1640
```

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 3060 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ix) FEATURE:
(A) NAME/KEY: CDS

```
          (B) LOCATION: 763..2202

     (xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

CCCACTTCCG GCGCTCGGGC TGCGCGAGGG CCCGTGCGAG CACATGATGG CGCTGCGGCT      60

CGTCCAGGTC CGGCACCGCG CCGAGCAGGA AGCACTGCGT CAGACCCCCG CGGGCCGCCA     120

GCTCATCCGC GCGGAGACGC GCTCCTACGT GCGGCGCGAG CCCTCCGGCC AGGAGCAGGT     180

GTACCGCGTC TCATTGGATG GGAAAGTGGT GGCGGTGGAG TGGGGCCCCC GCCAGGGGGA     240

GTCCCGCCGG CAGAAGCTCT GGTTCGACAC GGACGCCGAG GCGCGCACCG CCTACTTCAC     300

GCGCCTGGAG TCCTTGGCCG CGGAGGGATA TATCGATGCG GCTGCTTCAA TGATGTAGAA     360

CACGCAAGCC ACGGGGCCGC GGGCGCGCGG CGGAAAGGCA GGTGCGACGG AACGACAGAC     420

ACTCGTGCGA GCGACCGAGA GAGGTCCCAA GCCATCAGCC TCAGCGCCTC GAGCGCGAGA     480

GCGGCGTTGC GCCGCTCTGG TTGAATTGCA GGACACTCTC CGCAAGGTAG CCTGTTCTTG     540

GCTCTCTTCC CTCCGGTGAG TACCTCTCCG GCCGGGGAGC TGAACCAACG ACGCAACCGC     600

CGTTTCCCCG GCCGGAGAGG TACTCACCGG AGGGGAGAGC CGGTGAGGCT ACCGTGCCCC     660

AGGTGAGAAG GTGGTGCCTT CGGGCCTCCC TCGACCGCTC GCGCTCCGTC GCCCTGCCCT     720

GCCTCGCCCC CCCCACCTTG CTCACCGGCG CCAGGAGCCG TC ATG ACC GCC AAG        774
                                               Met Thr Ala Lys
                                                 1

CTG GAG TCA CAC GTC CCC GCC GCG CCC CCC GTC TCC GCC GAG GCG CCC       822
Leu Glu Ser His Val Pro Ala Ala Pro Pro Val Ser Ala Glu Ala Pro
  5                  10                  15                  20

GCC CCC ACC CGT CCC GAT GCC GCG AAG CAG GAG GCC CGC CGC GCC CAC       870
Ala Pro Thr Arg Pro Asp Ala Ala Lys Gln Glu Ala Arg Arg Ala His
                 25                  30                  35

CAC GAG GCG CTG CGC CTG CGG TGG AAG GCC ATC GAA GAG GCG GGC GGC       918
His Glu Ala Leu Arg Leu Arg Trp Lys Ala Ile Glu Glu Ala Gly Gly
             40                  45                  50

ACG GAC GCC TGG GTG CGG CAG CAG CTG GTG GCC AAG GGC GTC GCG GCG       966
Thr Asp Ala Trp Val Arg Gln Gln Leu Val Ala Lys Gly Val Ala Ala
         55                  60                  65

GAA GAG GTG GAC TTC GAG TCG CTC AGC GAC AAG CAG AAG GCG GCC TGG      1014
Glu Glu Val Asp Phe Glu Ser Leu Ser Asp Lys Gln Lys Ala Ala Trp
     70                  75                  80

AAG GAG AAG AAG AAG GCC GAG GCC ACC GAG CGG CGC GCG CAG AAG CGC      1062
```

```
Lys Glu Lys Lys Lys Ala Glu Ala Thr Glu Arg Arg Ala Gln Lys Arg
 85                  90                  95                 100

CTG GCG TGG GAG GCC TGG AAG GCC ACG CAC ATC CAC CAC CTG GGC GTG        1110
Leu Ala Trp Glu Ala Trp Lys Ala Thr His Ile His His Leu Gly Val
                105                 110                 115

GGG GTG CAC TGG GAC GAG GCC GGA GGG CCG GAC AAG TTC GAC GTG GCC        1158
Gly Val His Trp Asp Glu Ala Gly Gly Pro Asp Lys Phe Asp Val Ala
            120                 125                 130

GGG CGC GAG GAG CGG GCC AAG GCC AAC GGC TTG CCG GAG GGG TTG GAC        1206
Gly Arg Glu Glu Arg Ala Lys Ala Asn Gly Leu Pro Glu Gly Leu Asp
        135                 140                 145

TCG GTC GAG GCG CTG GCC AAA GCG CTG GGC ATC TCC GTG TCG CGC CTG        1254
Ser Val Glu Ala Leu Ala Lys Ala Leu Gly Ile Ser Val Ser Arg Leu
    150                 155                 160

CGC TGG TTC TCC TTC CAC CGC GAG GTG GAC ACG GGC ACG CAC TAC CAG        1302
Arg Trp Phe Ser Phe His Arg Glu Val Asp Thr Gly Thr His Tyr Gln
165                 170                 175                 180

ACG TGG GAG ATT CCG AAG CGG GAC GGC GGC AAG CGG ACG CTC ACC GCG        1350
Thr Trp Glu Ile Pro Lys Arg Asp Gly Gly Lys Arg Thr Leu Thr Ala
                185                 190                 195

CCG AAG CGG GAG CTC AAG GCC GTG CAG CGC TGG GTG CTC GCG AAC GTG        1398
Pro Lys Arg Glu Leu Lys Ala Val Gln Arg Trp Val Leu Ala Asn Val
            200                 205                 210

GTG GAG CGG CTG CCG GTG CAC GGG GCC GCG CAC GGC TTC GTG GCG GGG        1446
Val Glu Arg Leu Pro Val His Gly Ala Ala His Gly Phe Val Ala Gly
        215                 220                 225

CGC TCC ATC CTC ACC AAC GCG CTG GCC CAC CAG GGC GCG GAC GTG GTG        1494
Arg Ser Ile Leu Thr Asn Ala Leu Ala His Gln Gly Ala Asp Val Val
    230                 235                 240

GTG AAG GTG GAC ATG AAG GAC TTC TTC CCT TCC GTG ACG TGG CCC CGG        1542
Val Lys Val Asp Met Lys Asp Phe Phe Pro Ser Val Thr Trp Pro Arg
245                 250                 255                 260

GTC AAG GGA CTG CTG CGC AAG GGA GGA CTC CCG GAG AAC CTG GCG ACG        1590
Val Lys Gly Leu Leu Arg Lys Gly Gly Leu Pro Glu Asn Leu Ala Thr
                265                 270                 275

CTC CTG GCG CTG CTC TCC ACC GAG GCC CCG CGC GAG GTG GTG CGG TTC        1638
Leu Leu Ala Leu Leu Ser Thr Glu Ala Pro Arg Glu Val Val Arg Phe
            280                 285                 290

CGG GGA GAG ACG CTG TAC GTG GCC AAG GGC CCT CGC GCG CTG CCC CAG        1686
Arg Gly Glu Thr Leu Tyr Val Ala Lys Gly Pro Arg Ala Leu Pro Gln
        295                 300                 305
```

```
GGG GCC CCC ACC TCT CCG GCG CTG ACG AAC GCG CTG TGC CTG CGG CTG      1734
Gly Ala Pro Thr Ser Pro Ala Leu Thr Asn Ala Leu Cys Leu Arg Leu
    310                 315                 320

GAC AAG CGG CTC TCG GCG CTG TCG AAG CGG CTG GGC TTC ACG TAC ACG      1782
Asp Lys Arg Leu Ser Ala Leu Ser Lys Arg Leu Gly Phe Thr Tyr Thr
325                 330                 335                 340

CGC TAT GCG GAT GAC CTG ACG TTC TCC TGG CGG CGG GCG AAG AAG TCC      1830
Arg Tyr Ala Asp Asp Leu Thr Phe Ser Trp Arg Arg Ala Lys Lys Ser
                345                 350                 355

CGG CAG AAG GAA CTC CCC CTG GCG GAT GCG CCG GTG GCG CTG CTC CTG      1878
Arg Gln Lys Glu Leu Pro Leu Ala Asp Ala Pro Val Ala Leu Leu Leu
            360                 365                 370

GCG CGG GTG AAG GGT GTG CTG GAG GCC GAG GGT TTC ACG CTG CAC CCG      1926
Ala Arg Val Lys Gly Val Leu Glu Ala Glu Gly Phe Thr Leu His Pro
        375                 380                 385

GAC AAG ACG CGG GTG CAG CGC AAG GGC AGC CGG CAG CGG GTG ACG GGG      1974
Asp Lys Thr Arg Val Gln Arg Lys Gly Ser Arg Gln Arg Val Thr Gly
    390                 395                 400

CTC GTG GTG AAC GAG GCC CCC GAG GGC GTT CCG GGT GCC CGG GTG CCC      2022
Leu Val Val Asn Glu Ala Pro Glu Gly Val Pro Gly Ala Arg Val Pro
405                 410                 415                 420

CGC GAT GTG GTG CGG CGG CTG CGC GCG GCG ATC CAC AAC CGG GAG CAG      2070
Arg Asp Val Val Arg Arg Leu Arg Ala Ala Ile His Asn Arg Glu Gln
                425                 430                 435

GGC AAG CCC GGC CCC ACC GGG GAG ACG CTG GAG CAG CTC AAG GGG CTC      2118
Gly Lys Pro Gly Pro Thr Gly Glu Thr Leu Glu Gln Leu Lys Gly Leu
            440                 445                 450

GCG GCC TTC CTT CAC ATG ACG GAC GCG GAG AAG GGC CGC GCC TTC CTG      2166
Ala Ala Phe Leu His Met Thr Asp Ala Glu Lys Gly Arg Ala Phe Leu
        455                 460                 465

CGA CGG CTG GAG GCC CTC GAG AAG CGC CAG ACC GCC TGACCCTCAC           2212
Arg Arg Leu Glu Ala Leu Glu Lys Arg Gln Thr Ala
    470                 475                 480

TGGTCGTCCG GGGCATCGCA GCGGGCGCCG GGACGGACCG TCACCCCCCA GATCTCCATG    2272

CCATGCTGGG GATTCTGGGC GGTGAAGAAG ACTTCCCAGC CGAGACGGAC GAAGCCCTGC    2332

GGATCCGATG ACTCCTCGCC CGGGGCGATC TCCCGGAGGG GCACCGTTCC GACGTCCGTG    2392

CCATTGCTCA CCCAGGGCTC CCGGCCCCAG CCTTGGGTGT CCGCCGAGAA GAAGAGCAGC    2452

CCGGAGATGG CCGTCAGGTT CTCCGGCGAC GCATCCTCGG GGCCCGGCGC CAAATCCTTC    2512
```

```
AGCAGCAGGG TGCCCTTGGC GGTGCCATCG CTGGACCACA GCTCCCGGCC GTGGAGGCTG    2572

TCACTCGCGG CGAAGTAGAG CATCCCATTC AGCGCCTTGA TGGCGCTGGG CGCCGAGCTG    2632

TCCGGACCCG GCCAGATGTC CTTCACCCGG ACCGTGCCAT GCGACGTGCC ATCGCTGACC    2692

CACAGCTCCT CGCCCTCGGG CTGGCCCCAG AACTCGGGCT CGCCTCCCCC GGCGCTGAAG    2752

AAGATCTTCC CCCCGAGCGC CGTGAGATCA TGCGGATAGA GGCCGGGGAA GAAGCGCAGC    2812

TGCTCGGAGA CGGTGCCTCT GGAGCACCAC AGGCTGGCCT CGCCTTCGTC ATTGTCGAGC    2872

AGGAAGAAGA GCACCGAGTC CGCCGCGGTG AACGCGGAGA GGAAGTTGTC CTCGGGGCCC    2932

GTGAAGACAG ACGTGGTGCT GGACAGCCCC AGGCTGCGCC AGATGAACAC CTCGTCATTG    2992

ACGTTGGCCA CGAAGAAGAG CGCATCGCCG ACCCGGGTGA GCCGGCGCGG GCTGGAGCTG    3052

CCGGGCAC                                                             3060
```

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 2788 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 2..103

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 707..1654

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 1644..2591

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

```
T TTC GAG AAG CGC CAT ACC AAA CAG GGG ATA CAG ACC AAC CTG ACG          46
  Phe Glu Lys Arg His Thr Lys Gln Gly Ile Gln Thr Asn Leu Thr
    1               5                  10                  15

CTG AAA GAG GAA AGC TAC GGC GAC TGG CTG CCG AAG TGC GAC GAC CCC        94
Leu Lys Glu Glu Ser Tyr Gly Asp Trp Leu Pro Lys Cys Asp Asp Pro
                20                  25                  30

GCA GCA ACA TAACCTCACT CAGACCGGCA ACAGCCGGTC TTTTCCTTTC               143
```

```
Ala Ala Thr

TGGCCATTGC CACAAGGTGA ACAATCCACT GTTCACCCTT CACCGTTTAT TCACCCTTTA      203

TCACTATGAA ATTATTAATA AAAAACCAGA GGTGAACAGT GTGAACAGTA AAACCTGAAA      263

AAACTTTTTA TCACCCCGCG CATCGCCCGA CTGGACAGAT CCAGAACGAG CAAAAATCAC      323

AAAGGTGACG AGTCGACTGT TCACTCTTCA CCAACTCATC ACCACCTAAC CACATGATAT      383

AAAATGATAA ATAATCGAGG TGAACAGTTA AATGCAAAAA AACTTTTTCT CAGCTCTTGG      443

ATAAAAGAAA ATTAATTCAC ATCAATAGCT TTCCTCTTGA ATCCTCTTGA GGTTTATGAG      503

AGCGTAACAG AGCCAAACCT AGCATTTTAT GGGTTAATAG CCCATCGCGC ATGAGTCATG      563

GTTTCGCCTA GTATTTTAGC TATGCCCGTC GTTCAGTTCG CTGAGCGGCG GCTGGGGGCC      623

ACCGATCAGC GAACTGATCG ACGTGCTCAA GTAGGTTTGG CTCTTTTAGT CCTCTACCAT      683

CAAGGTGCAT AAGGATATTC TCG ATG CTG ACT CAG CTA AAA AAA AAT GGT         733
                          Met Leu Thr Gln Leu Lys Lys Asn Gly
                           1               5

ACT GAG GTA TCT AGA GCA ACC GCG TTA TTT TCA TCA TTC GTT GAA AAG      781
Thr Glu Val Ser Arg Ala Thr Ala Leu Phe Ser Ser Phe Val Glu Lys
 10                 15                  20                  25

AAC AAA GTA AAA TGT CCT GGT AAT GTA AAA AAA TTC GTC TTT CTG TGT      829
Asn Lys Val Lys Cys Pro Gly Asn Val Lys Lys Phe Val Phe Leu Cys
                 30                  35                  40

GGT GCT AAC AAA AAC AAT GGA GAA CCA TCA GCA AGA CGA TTG GAA TTA      877
Gly Ala Asn Lys Asn Asn Gly Glu Pro Ser Ala Arg Arg Leu Glu Leu
             45                  50                  55

ATA AAT TTT TCT GAA AGG TAT TTG AAT AAC TGT CAC TTT TTT CTT GCT      925
Ile Asn Phe Ser Glu Arg Tyr Leu Asn Asn Cys His Phe Phe Leu Ala
         60                  65                  70

GAA CTA GTT TTC AAA GAA TTA AGC ACC GAT GAA GAA TCA TTA TCT GAT      973
Glu Leu Val Phe Lys Glu Leu Ser Thr Asp Glu Glu Ser Leu Ser Asp
     75                  80                  85

AAT TTA TTA GAT ATC GAA GCT GAC TTA TCT AAA TTA GCT GAT CAT ATT     1021
Asn Leu Leu Asp Ile Glu Ala Asp Leu Ser Lys Leu Ala Asp His Ile
 90                  95                 100                 105

ATC ATT GTT TTA GAA AGT TAT TCA TCT TTC ACG GAA CTT GGT GCA TTC     1069
Ile Ile Val Leu Glu Ser Tyr Ser Ser Phe Thr Glu Leu Gly Ala Phe
                110                 115                 120

GCA TAC AGC AAG CAA TTA CGC AAG AAA TTA ATA ATA GTT AAC AAT ACA     1117
```

```
Ala Tyr Ser Lys Gln Leu Arg Lys Lys Leu Ile Ile Val Asn Asn Thr
            125                 130                 135

AAA TTT ATA AAT GAG AAA TCA TTT ATA AAT ATG GGA CCA ATA AAG GCT      1165
Lys Phe Ile Asn Glu Lys Ser Phe Ile Asn Met Gly Pro Ile Lys Ala
        140                 145                 150

ATT ACT CAG CAA TCA CAA CAA TCT GGT CAT TTC TTA CAT TAT AAA ATG      1213
Ile Thr Gln Gln Ser Gln Gln Ser Gly His Phe Leu His Tyr Lys Met
    155                 160                 165

ACA GAA GGT ATT GAA AGT ATA GAG CGC TCT GAT GGG ATT GGC GAA ATA      1261
Thr Glu Gly Ile Glu Ser Ile Glu Arg Ser Asp Gly Ile Gly Glu Ile
170                 175                 180                 185

TTC GAC CCC CTA TAT GAT ATT CTT TCT AAG AAC GAC AGA GCA ATT TCA      1309
Phe Asp Pro Leu Tyr Asp Ile Leu Ser Lys Asn Asp Arg Ala Ile Ser
                190                 195                 200

AGA ACT TTA AAA AAA GAA GAG TTA GAT CCT TCC AGT AAC TTC AAT AAA      1357
Arg Thr Leu Lys Lys Glu Glu Leu Asp Pro Ser Ser Asn Phe Asn Lys
            205                 210                 215

GAC TCA GTA CGA TTT ATT CAT GAC GTA ATT TTT GTA TGT GGT CCT TTG      1405
Asp Ser Val Arg Phe Ile His Asp Val Ile Phe Val Cys Gly Pro Leu
        220                 225                 230

CAA CTT AAT GAA CTC ATC GAA ATA ATC ACA AAA ATA TTT GGC ACA GAA      1453
Gln Leu Asn Glu Leu Ile Glu Ile Ile Thr Lys Ile Phe Gly Thr Glu
    235                 240                 245

AGC CAT TAC AAA AAA AAT CTT CTA AAG CAC CTT GGT ATT CTA ATA GCT      1501
Ser His Tyr Lys Lys Asn Leu Leu Lys His Leu Gly Ile Leu Ile Ala
250                 255                 260                 265

ATT AGA ATA ATA TCA TGC ACA AAT GGG ATT TAT TAT TCT TTG TAT AAA      1549
Ile Arg Ile Ile Ser Cys Thr Asn Gly Ile Tyr Tyr Ser Leu Tyr Lys
                270                 275                 280

GAA TAT TAT TTT AAA TAT GAC TTT GAC ATT GAC AAC ATA TCA TCA ATG      1597
Glu Tyr Tyr Phe Lys Tyr Asp Phe Asp Ile Asp Asn Ile Ser Ser Met
            285                 290                 295

TTT AAA GTT TTT TTC CTC AAG AAC AAG CCA GAA AGG ATG AGG GTA TAT      1645
Phe Lys Val Phe Phe Leu Lys Asn Lys Pro Glu Arg Met Arg Val Tyr
        300                 305                 310

GAG AAT ATA TAGCCTAATT GATTCTCAGA CATTGATGAC TAAGGGATTT              1694
Glu Asn Ile
    315

GCTTCTGAAG TAATGCGATC ACCTGAGCCG CCAAAAAAAT GGGATATAGC TAAGAAAAAA    1754

GGAGGTATGA GAACAATTTA TCACCCGTCA TCAAAAGTTA AATTAATTCA ATATTGGTTA    1814
```

```
ATGAATAATG TTTTTTCGAA GCTCCCAATG CATAATGCTG CATATGCATT TGTTAAAAAC    1874

CGATCAATAA AAAGCAATGC TTTATTACAT GCCGAATCAA AGAATAAGTA TTATGTGAAA    1934

ATAGATCTCA AAGATTTTTT CCCTTCAATA AAATTTACTG ATTTTGAGTA CGCATTCACT    1994

CGTTATCGAG ATCGCATTGA ATTTACTACA GAATATGATA AGGAGTTACT ACAACTTATA    2054

AAAACGATCT GCTTTATATC AGATAGCACT CTCCCTATCG GGTTTCCTAC ATCTCCATTA    2114

ATTGCAAACT TTGTGGCAAG AGAACTTGAT GAAAAACTGA CGCAAAAACT AAATGCAATT    2174

GATAAACTTA ATGCCACTTA TACACGATAT GCTGATGATA TTATTGTCTC TACAAATATG    2234

AAAGGGGCTA GCAAATTAAT TCTGGATTGT TTTAAAAGAA CAATGAAAGA GATTGGTCCA    2294

GACTTTAAAA TTAACATTAA AAAATTTAAG ATTTGTAGTG CTTCGGGAGG AAGTATAGTA    2354

GTTACCGGAT TGAAAGTTTG CCACGATTTT CATATTACAT TACATAGATC AATGAAAGAT    2414

AAAATAAGAT TGCATCTTTC TCTTTTATCA AAGGGCATAT TAAAAGATGA AGATCATAAT    2474

AAACTTTCTG GTTATATTGC TTATGCAAAA GATATAGACC CTCATTTTTA TACAAAACTG    2534

AACAGAAAAT ATTTTCAAGA AATAAAATGG ATTCAGAATC TCCACAACAA AGTTGAATAA    2594

ACTTTATATT TTGGATGCAC CCCAATAACT TCATTGATTA AATTGGGAAC AATATAGGCT    2654

TTTCAGGATG ACCTACACTC TAGAGAATGT GTATACAAAA GTGTATAAGT TATTTTCAAA    2714

CCTATATAAA ATACAGCAAA ATCAATGCAT TGGCGGCATT TTACCACTCC TGTGATCTTC    2774

CGCCAAAATG CCTC                                                      2788
```

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 316 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

```
Met Arg Ile Tyr Ser Leu Ile Asp Ser Gln Thr Leu Met Thr Lys Gly
  1               5                  10                  15

Phe Ala Ser Glu Val Met Arg Ser Pro Glu Pro Pro Lys Lys Trp Asp
             20                  25                  30

Ile Ala Lys Lys Lys Gly Gly Met Arg Thr Ile Tyr His Pro Ser Ser
         35                  40                  45
```

```
Lys Val Lys Leu Ile Gln Tyr Trp Leu Met Asn Asn Val Phe Ser Lys
     50                  55                  60

Leu Pro Met His Asn Ala Ala Tyr Ala Phe Val Lys Asn Arg Ser Ile
 65                  70                  75                  80

Lys Ser Asn Ala Leu Leu His Ala Glu Ser Lys Asn Lys Tyr Tyr Val
                 85                  90                  95

Lys Ile Asp Leu Lys Asp Phe Phe Pro Ser Ile Lys Phe Thr Asp Phe
            100                 105                 110

Glu Tyr Ala Phe Thr Arg Tyr Arg Asp Arg Ile Glu Phe Thr Thr Glu
        115                 120                 125

Tyr Asp Lys Glu Leu Leu Gln Leu Ile Lys Thr Ile Cys Phe Ile Ser
    130                 135                 140

Asp Ser Thr Leu Pro Ile Gly Phe Pro Thr Ser Pro Leu Ile Ala Asn
145                 150                 155                 160

Phe Val Ala Arg Glu Leu Asp Glu Lys Leu Thr Gln Lys Leu Asn Ala
                165                 170                 175

Ile Asp Lys Leu Asn Ala Thr Tyr Thr Arg Tyr Ala Asp Asp Ile Ile
            180                 185                 190

Val Ser Thr Asn Met Lys Gly Ala Ser Lys Leu Ile Leu Asp Cys Phe
        195                 200                 205

Lys Arg Thr Met Lys Glu Ile Gly Pro Asp Phe Lys Ile Asn Ile Lys
    210                 215                 220

Lys Phe Lys Ile Cys Ser Ala Ser Gly Gly Ser Ile Val Val Thr Gly
225                 230                 235                 240

Leu Lys Val Cys His Asp Phe His Ile Thr Leu His Arg Ser Met Lys
                245                 250                 255

Asp Lys Ile Arg Leu His Leu Ser Leu Leu Ser Lys Gly Ile Leu Lys
            260                 265                 270

Asp Glu Asp His Asn Lys Leu Ser Gly Tyr Ile Ala Tyr Ala Lys Asp
        275                 280                 285

Ile Asp Pro His Phe Tyr Thr Lys Leu Asn Arg Lys Tyr Phe Gln Glu
    290                 295                 300

Ile Lys Trp Ile Gln Asn Leu His Asn Lys Val Glu
305                 310                 315
```

(2) INFORMATION FOR SEQ ID NO:41:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 1602 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 548..1507

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

```
TGGCATCTAT TAAGAAGGTT AGGAAAGAAA ATAAAGTATC AAAAGATATT GGAAATATAT      60

TATACGCAGA GCGTTTCTAT TGCCTTGTAT CTATTTACTG GATAGTGTCA ACTACCGCAC     120

ACTGTGTGAA CTAGCTTTTA AAGCGATAAA GCAAGATGAT GTTTTATCTA AAATTATTGT     180

TAGATCCGTT GTTTCTCGTC TAATAAATGA ACGAAAAATA CTTCAAATGA CTGATGGTTA     240

TCAGGTCACT GCTTTGGGGG CTAGCTATGT TAGGAGCGTC TTTGATAGAA AGACACTTGA     300

CCGATTGCGG CTTGAGATTA TGAATTTTGA AAACCGTAGA AAATCAACAT TTAACTATGA     360

TAAGATTCCG TATGCGCACC CTTAGCGAGA GGTTTATCAT TAAGGTCAAC CTCTGGATGT     420

TGTTTCGGCA TCCTGCATTG AATCTGAGTT ACTGTCTGTT TTCCTTGTTG GAACGGAGAG     480

CATCGCCTGA TGCTCTCCGA GCCAACCAGG AAACCCGTTT TTTCTGACGT AAGGGTGCGC     540

AACTTTC ATG AAA TCC GCT GAA TAT TTG AAC ACT TTT AGA TTG AGA AAT       589
        Met Lys Ser Ala Glu Tyr Leu Asn Thr Phe Arg Leu Arg Asn
         1               5                  10

CTC GGC CTA CCT GTC ATG AAC AAT TTG CAT GAC ATG TCT AAG GCG ACT       637
Leu Gly Leu Pro Val Met Asn Asn Leu His Asp Met Ser Lys Ala Thr
 15                  20                  25                  30

CGC ATA TCT GTT GAA ACA CTT CGG TTG TTA ATC TAT ACA GCT GAT TTT       685
Arg Ile Ser Val Glu Thr Leu Arg Leu Leu Ile Tyr Thr Ala Asp Phe
                 35                  40                  45

CGC TAT AGG ATC TAC ACT GTA GAA AAG AAA GGC CCA GAG AAG AGA ATG       733
Arg Tyr Arg Ile Tyr Thr Val Glu Lys Lys Gly Pro Glu Lys Arg Met
             50                  55                  60

AGA ACC ATT TAC CAA CCT TCT CGA GAA CTT AAA GCC TTA CAA GGA TGG       781
Arg Thr Ile Tyr Gln Pro Ser Arg Glu Leu Lys Ala Leu Gln Gly Trp
         65                  70                  75

GTT CTA CGT AAC ATT TTA GAT AAA CTG TCG TCA TCT CCT TTT TCT ATT       829
Val Leu Arg Asn Ile Leu Asp Lys Leu Ser Ser Ser Pro Phe Ser Ile
     80                  85                  90
```

```
GGA TTT GAA AAG CAC CAA TCT ATT TTG AAT AAT GCT ACC CCG CAT ATT          877
Gly Phe Glu Lys His Gln Ser Ile Leu Asn Asn Ala Thr Pro His Ile
 95                 100                 105                 110

GGG GCA AAC TTT ATA CTG AAT ATT GAT TTG GAG GAT TTT TTC CCA AGT          925
Gly Ala Asn Phe Ile Leu Asn Ile Asp Leu Glu Asp Phe Phe Pro Ser
                115                 120                 125

TTA ACT GCT AAC AAA GTT TTT GGA GTG TTC CAT TCT CTT GGT TAT AAT          973
Leu Thr Ala Asn Lys Val Phe Gly Val Phe His Ser Leu Gly Tyr Asn
            130                 135                 140

CGA CTA ATA TCT TCA GTT TTG ACA AAA ATA TGT TGT TAT AAA AAT CTG         1021
Arg Leu Ile Ser Ser Val Leu Thr Lys Ile Cys Cys Tyr Lys Asn Leu
        145                 150                 155

CTA CCA CAA GGT GCT CCA TCA TCA CCT AAA TTA GCT AAT CTA ATA TGT         1069
Leu Pro Gln Gly Ala Pro Ser Ser Pro Lys Leu Ala Asn Leu Ile Cys
    160                 165                 170

TCT AAA CTT GAT TAT CGT ATT CAG GGT TAT GCA GGT AGT CGG GGC TTG         1117
Ser Lys Leu Asp Tyr Arg Ile Gln Gly Tyr Ala Gly Ser Arg Gly Leu
175                 180                 185                 190

ATA TAT ACG AGA TAT GCC GAT GAT CTC ACC TTA TCT GCA CAG TCT ATG         1165
Ile Tyr Thr Arg Tyr Ala Asp Asp Leu Thr Leu Ser Ala Gln Ser Met
                195                 200                 205

AAA AAG GTT GTT AAA GCA CGT GAT TTT TTA TTT TCT ATA ATC CCA AGT         1213
Lys Lys Val Val Lys Ala Arg Asp Phe Leu Phe Ser Ile Ile Pro Ser
            210                 215                 220

GAA GGA TTG GTT ATT AAC TCA AAA AAA ACT TGT ATT AGT GGG CCT CGT         1261
Glu Gly Leu Val Ile Asn Ser Lys Lys Thr Cys Ile Ser Gly Pro Arg
        225                 230                 235

AGT CAG AGG AAA GTT ACA GGT TTA GTT ATT TCA CAA GAG AAA GTT GGG         1309
Ser Gln Arg Lys Val Thr Gly Leu Val Ile Ser Gln Glu Lys Val Gly
    240                 245                 250

ATA GGT AGA GAA AAA TAT AAA GAA ATT AGA GCA AAG ATA CAT CAT ATA         1357
Ile Gly Arg Glu Lys Tyr Lys Glu Ile Arg Ala Lys Ile His His Ile
255                 260                 265                 270

TTT TGC GGT AAG TCT TCT GAG ATA GAA CAC GTT AGG GGA TGG TTG TCA         1405
Phe Cys Gly Lys Ser Ser Glu Ile Glu His Val Arg Gly Trp Leu Ser
                275                 280                 285

TTT ATT TTA AGT GTG GAT TCA AAA AGC CAT AGG AGA TTA ATA ACT TAT         1453
Phe Ile Leu Ser Val Asp Ser Lys Ser His Arg Arg Leu Ile Thr Tyr
            290                 295                 300

ATT AGC AAA TTA GAA AAA AAA TAT GGA AAG AAC CCT TTA AAT AAA GCG         1501
Ile Ser Lys Leu Glu Lys Lys Tyr Gly Lys Asn Pro Leu Asn Lys Ala
```

```
       305                 310                 315

AAG ACC TAATGGTCTT CGTTTTAAAA CTAAAGCTCA TAGGTTGAAA AATTGAGCAC       1557
Lys Thr
    320

TTCTTCGTCC AACCAGTTAT TTAGTTCCTG CAATCGTTTC TGCAG                    1602
```

(2) INFORMATION FOR SEQ ID NO:42:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 1540 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: double
(D) TOPOLOGY: linear

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 396..1352

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:

```
TCACCCTGAA AGACCTGATT GCTTACCTGG AAGAGAAGCC GGAAATGGCG GAACATCTGG      60

CGGCGGTTAA GGCCTATCGC GAAGAGTTCG GCGTTTAAAA ATATGCGCTG TGCAGGGTTT     120

TTGCTGTGCG CAGCGTGATG CGCTTCAAGA TATCGTGTTA ATCTGCTTTC GCCAGCAGTG     180

GCAATAGCGT TTCCGGCCTT TTGTGCCGGG AGGGTCGGCG AGTCGCTGAC TTAACGCCAG     240

TAGTATGTCC ATATACCCAA AGTCGCTTCA TTGTACCTGA GTACGCTTCG CGTACGTCGC     300

GCTGACGCGC TCAGTACAGT TACGCGCCTT CGGGATGGTT TAATGGTATT GCCGCTGTTG     360

GCGCCTCTTT TGGCCGCCGT GATGTGGAGA GTGGA ATG GAT GCT ACC CGG ACA        413
                                      Met Asp Ala Thr Arg Thr
                                       1               5

ACC CTT CTG GCG CTC GAT TTG TTC GGC TCG CCG GGC TGG AGC GCC GAT      461
Thr Leu Leu Ala Leu Asp Leu Phe Gly Ser Pro Gly Trp Ser Ala Asp
            10                  15                  20

AAA GAA ATA CAG CGA CTG CAT GCG CTC AGT AAT CAT GCC GGA CGC CAT      509
Lys Glu Ile Gln Arg Leu His Ala Leu Ser Asn His Ala Gly Arg His
        25                  30                  35

TAC CGA CGC ATT ATT CTT TCT AAA CGC CAC GGT GGT CAG CGG CTG GTG      557
Tyr Arg Arg Ile Ile Leu Ser Lys Arg His Gly Gly Gln Arg Leu Val
    40                  45                  50

TTA GCC CCT GAT TAC TTG CTC AAA ACC GTA CAG CGC AAC ATT CTT AAG      605
```

```
Leu Ala Pro Asp Tyr Leu Leu Lys Thr Val Gln Arg Asn Ile Leu Lys
 55                  60                  65                  70

AAC GTC CTT TCA CAA TTT CCG CTT TCC CCT TTT GCT ACA GCC TAC CGA      653
Asn Val Leu Ser Gln Phe Pro Leu Ser Pro Phe Ala Thr Ala Tyr Arg
                 75                  80                  85

CCA GGT TGC CCA ATC GTC AGC AAC GCG CAG CCA CAC TGC CAA CAG CCG      701
Pro Gly Cys Pro Ile Val Ser Asn Ala Gln Pro His Cys Gln Gln Pro
             90                  95                 100

CAG ATC CTG AAA CTC GAT ATC GAA AAC TTT TTC GAT AGC ATT AGC TGG      749
Gln Ile Leu Lys Leu Asp Ile Glu Asn Phe Phe Asp Ser Ile Ser Trp
        105                 110                 115

TTA CAG GTC TGG CGT GTG TTT CGC CAG GCC CAG TTG CCA CGT AAT GTG      797
Leu Gln Val Trp Arg Val Phe Arg Gln Ala Gln Leu Pro Arg Asn Val
    120                 125                 130

GTA ACC ATG CTG ACC TGG ATT TGT TGT TAT AAC GAC GCG TTA CCG CAG      845
Val Thr Met Leu Thr Trp Ile Cys Cys Tyr Asn Asp Ala Leu Pro Gln
135                 140                 145                 150

GGG GCA CCA ACT TCG CCA GCC ATT TCC AAT CTT GTG ATG CGC CGT TTT      893
Gly Ala Pro Thr Ser Pro Ala Ile Ser Asn Leu Val Met Arg Arg Phe
                155                 160                 165

GAT GAA CGC ATA GGG GAA TGG TGT CAG GCT CGG GGA ATT ACC TAC ACC      941
Asp Glu Arg Ile Gly Glu Trp Cys Gln Ala Arg Gly Ile Thr Tyr Thr
            170                 175                 180

CGC TAC TGC GAT GAC ATG ACC TTT TCA GGT CAC TTC AAT GCC CGC CAG      989
Arg Tyr Cys Asp Asp Met Thr Phe Ser Gly His Phe Asn Ala Arg Gln
        185                 190                 195

GTT AAA AAT AAA GTG TGC GGA TTG TTA GCG GAG CTG GGC CTG AGC CTC     1037
Val Lys Asn Lys Val Cys Gly Leu Leu Ala Glu Leu Gly Leu Ser Leu
    200                 205                 210

AAT AAA CGC AAA GGC TGC CTG ATA GCT GCC TGT AAG CGC CAG CAA GTA     1085
Asn Lys Arg Lys Gly Cys Leu Ile Ala Ala Cys Lys Arg Gln Gln Val
215                 220                 225                 230

ACC GGG ATT GTT GTT AAT CAC AAG CCA CAG CTT GCC CGT GAA GCG CGC     1133
Thr Gly Ile Val Val Asn His Lys Pro Gln Leu Ala Arg Glu Ala Arg
                235                 240                 245

CGG GCG CTG CGT CAG GAG GTG CAT TTG TGC CAA AAA TAT GGC GTT ATT     1181
Arg Ala Leu Arg Gln Glu Val His Leu Cys Gln Lys Tyr Gly Val Ile
            250                 255                 260

TCG CAT CTT AGT CAT CGT GGT GAA CTT GAT CCT TCT GGC GAT CTC CAC     1229
Ser His Leu Ser His Arg Gly Glu Leu Asp Pro Ser Gly Asp Leu His
        265                 270                 275
```

```
GCA CAG GCA ACG GCG TAT CTT TAT GCT TTG CAG GGA AGA ATA AAC TGG        1277
Ala Gln Ala Thr Ala Tyr Leu Tyr Ala Leu Gln Gly Arg Ile Asn Trp
    280                 285                 290

TTA TTG CAA ATC AAC CCT GAG GAT GAG GCC TTT CAA CAG GCG AGA GAG        1325
Leu Leu Gln Ile Asn Pro Glu Asp Glu Ala Phe Gln Gln Ala Arg Glu
295                 300                 305                 310

AGT GTA AAG CGA ATG CTG GTT GCA TGG TAAGAAAAGC GTCAGGCAGA              1372
Ser Val Lys Arg Met Leu Val Ala Trp
                315

CGTTTCTGCC TGACCGTTTA GGGGAGAATT ACTGCAACTG CGCGGCAATT AGCGGCCAGC      1432

GGGCGTCAAA ATCATCCGTC GGGCGGTATT TAAACTCGCT GCGGACAAAA CGTGACAGCA      1492

TACCTTCACA GAAGGCCAGG ATCTGGCTTG CCAGCAGGGT TTCATCGG                   1540
```

(2) INFORMATION FOR SEQ ID NO:43:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 4 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

```
Tyr Xaa Asp Asp
 1          4
```

(2) INFORMATION FOR SEQ ID NO:44:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 4 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

```
Ser Xaa Xaa Xaa
 1          4
```

(2) INFORMATION FOR SEQ ID NO:45:

Sub G' concluded

(i) SEQUENCE CHARACTERISTICS:
    (A) LENGTH: 4 amino acids
    (B) TYPE: amino acid
    (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:45:

```
Xaa Val Thr Gly
  1           4
```

United States Patent & Trademark Office

Office of Initial Patent Examination -- Scanning Division

Application deficiencies found during scanning:

□ Page(s)___________ of ________________________________ were not present for scanning.
(Document title)

□ Page(s)___________ of ________________________________ were not present for scanning.
(Document title)

☑ ***Scanned copy is best available.*** Drawings are dark, and there are lines in specification and sequence